A critical study of the use of an evaluation model designed to measure the outcomes and impact of development interventions

Submitted by: Tracey Phillips (Student Number: wcktra001)

A minor dissertation submitted in partial fulfillment of the requirements for the award of the degree of MPhil: Sociology

Faculty of the Humanities

University of Cape Town

September 2013

COMPULSORY DECLARATION

This work has not been previously submitted in whole, or in part, for the award of any degree. It is my own work. Each significant contribution to, and quotation in, this dissertation from the work, or works, of other people has been attributed, and has been cited and referenced.

Signature:_____ Date:_____

Acknowledgements

This research project was made possible thanks to the many people who gave so generously of their time and input.

I would like to thank my supervisor, Dr Jacques de Wet, for his input, motivation and guidance throughout the research process. It has been a valuable learning experience – and one which will stand me in good stead in future research endeavours.

I would also like to thank those client organisation representatives who agreed to participate in this study, and for doing so in such an honest, informative and engaged manner. Their insights and experiences contributed enormously to this report and offered a wealth of information on evaluation processes and approaches as well as client-consultancy interaction.

I would like to convey a special word of thanks to the four partners of Southern Hemisphere - for giving me the opportunity to undertake a study of this nature and for allowing me to interrogate their evaluation model in a critical and honest manner. I would also like to thank them for taking the time to meet with me for interviews and for their invaluable, on-going support and endorsement of this project in the interests of reflection and learning.

Lastly, I would like to thank my friends and family for their continuous encouragement and unswerving faith in my ability to undertake this journey. It would not have been possible without you.

Abstract

This study investigates a naturalistic evaluation model's ability to assess the outcomes and impact of development interventions in a rigorous manner. The study was undertaken by means of a meta-evaluation of five evaluation projects conducted by a socio-economic development consultancy situated in Cape Town. This meta-evaluation process was based upon four evaluation quality or 'trustworthiness' criteria proposed by Guba and Lincoln (1989); namely, credibility, transferability, dependability and confirmability. These four criteria were conceptualised, operationalised and applied to the evaluation projects under review.

Qualitative research methods were utilised, including an extensive project document review coupled with in-depth, semi-structured interviews, which were conducted face-to-face, telephonically or via Skype. The interview respondents included four evaluators / consultancy staff and seven representatives from the relevant client organisations.

Data analysis was undertaken using interview coding reports generated via the use of the software package, NVivo 10, in conjunction with an Excel spreadsheet meta-evaluation summary, compiled following an extensive project document review. Key themes, relating to each of the four quality criteria, that emerged from the interview data were cross-checked against the same themes in the meta-evaluation summary. Conclusions were drawn for each criterion based on areas of convergence and divergence between the data sources.

The findings of this study indicate that the evaluation model under review satisfies the criteria of credibility and transferability, but only partially meets the criteria of dependability and confirmability. This is predominantly due to a lack of methodological consistency, clarity and transparency.

The study concludes that outcome and impact assessments can be rigorously assessed through the use of an evaluation model that is situated within the naturalistic paradigm. However, it is proposed a) that evaluation studies be approached from a pragmatic perspective as opposed to adhering to evaluation processes that are paradigm-specific, and b) that all evaluation processes should be underpinned by appropriate standards of rigorous practice to ensure results that are both relevant and credible, irrespective of one's philosophical approach to evaluation research methods.

List of Acronyms

ACES	AIDS Community Educators
DAC	Development Assistance Committee
DRC	Democratic Republic of the Congo
HAICU	HIV / AIDS Institutional Coordination Unit
IOM	International Organization for Migration
M&E	Monitoring and Evaluation
MDGs	Millennium Development Goals
MIT	Massachusetts Institute of Technology
NGO	Non-governmental organisation
PDT	Ponahalo De Beers Trust
RCT	Randomised control trial
SH	Southern Hemisphere
UCT	University of Cape Town
UK	United Kingdom
UNICEF	United Nations Children's Fund
UNJPHT	United Nations Joint Programme on Human Trafficking
US	United States

Index

Chapter 1 Introduction

1.1 Introduction to the study

With a growing scarcity of available funding and an increasing emphasis on accountability and learning, it is now widely accepted within the development sector that implementing a development programme or project is insufficient; one also needs to establish (and clearly communicate or possess the ability to indicate) the effectiveness thereof. This is mirrored in the move away from traditional implementation and output-focused evaluation models towards results-based evaluation models; that is, a move away from narrowly investigating and noting the inputs, activities and outputs of a development intervention, towards a greater emphasis on measurement of results; which include the outcomes and possible impact of such an intervention. The rationale behind such a move towards results-based evaluation models is obvious. If one looks at outcomes and possible impacts or changes for the target group/s and beneficiaries of a programme, this will allow for a greater depth of insight into the intervention's strengths and weaknesses, plus it will allow for evidence-based decision-making in terms of future interventions and/or the potential 'going to scale' of an effective development programme. In turn, enhanced levels of intervention effectiveness will be encouraged, together with heightened levels of credibility for project implementers, who would seek such status in their attempt to procure funding for future projects.

Whilst the reasons for conducting results-based evaluations are clear; the defining and measuring of development intervention outcomes – and, in particular, impact - is not. Not only is the conceptualisation of the term 'impact' widely debated within development studies literature; so too are the evaluation methods employed to measure and report upon it. Furthermore, many evaluation theorists and practitioners, particularly those positioned within a positivist / post-positivist paradigm, argue that the 'quality' and 'rigour' of many outcome and impact evaluations - particularly those employing qualitative methods - are not subject to sufficient levels of critical assessment and reflection. Counter-arguments from those working in the field of naturalistic / interpretivist inquiry range from a rejection of the use of quality assessments and criteria for qualitative research (Smith 1984, 1990 cited in Spencer et al) to the suggestion of formulating quality criteria common to both naturalistic and positivist inquiry (Morse et al 2002, Le Compte and Goetz 1982, Kirk and Miller 1986). Mid-way between these two extremes lie those who propose the development and application of a set of 'alternative' criteria deemed more appropriate for assessing the rigour of naturalistic / interpretivist evaluation research (Guba and Lincoln 1989, 2001).

Evaluation literature notes that the field is currently characterised by a lack of research that focuses specifically upon a) practical methods and activities employed in evaluation studies, and b) the quality of the findings of such evaluations. Mark (2008: 118 - 119) states that the field is currently "…ripe for research…" particularly in terms of examining the components or 'parts' of a given 'model' of evaluation, and the effect of evaluation processes upon an evaluation's results. It is this 'gap' in evaluation literature that this research project aims to address; namely, the lack of critical study of - and reflection on - evaluation methods, practices and processes; coupled with appropriate and informative assessment of the quality of subsequent evaluation results. Hence, this study outlines a rigour and quality assessment of the process and products of a specific evaluation model which operates within a naturalistic / interpretivist paradigm. In this way, it hopes to make a contribution towards the assessment - and possible improvement - of the model whilst simultaneously examining possible means of quality assessment, and assurance, of naturalistic evaluation research.

1.2 Purpose and scope of the study

This research report outlines a meta-evaluation of five outcome and / or impact evaluations, conducted over the course of the past four years by a Cape Town-based socio-economic development consultancy, namely Southern Hemisphere (SH). These five evaluations are as follows:

Table 1: Southern Hemisphere evaluation projects under review

Evaluation Project	Date / time period of study
The Research Study of Student Risk Behaviour and Evaluation of ACES Programme for HAICU	March 2009 – July 2009
Save the Children UK's Response to the Situation in Musina since 2008	July 2010 – September 2010
The Outcome Assessment of the LEGO Care for Education Project: "Developing Talents through Creative Play" in Atteridgeville Township	September 2010 – November 2011
The Evaluation of the Ponahalo De Beers Trust (PDT) Programme	September 2011 – April 2012
The Evaluation of the United Nations Joint Programme on Human Trafficking (UNJPHT)	August 2012 – December 2012

The selection of the above evaluation projects was based upon two key criteria, namely a) all of the research projects had to include either an impact or an outcomes assessment, thus making them relevant to the central research question; and b) the research projects had to be undertaken within the past four years to ensure a sufficient level of institutional memory

- on the part of the consultancy partners as well as the client organisations - to allow for the collection of 'rich' and in-depth data.

The unit of analysis or focus of this research project will be the particular evaluation process or 'model' employed by the consultancy[1], which will be further outlined in the sections that follow. The ability of this evaluation model to deliver a rigorous evaluation - of a development intervention's outcomes and/or impact - will be interrogated via a meta-evaluation process, utilising a set of 'philosophically appropriate' research quality criteria, which are based upon the work of Lincoln and Guba (1981, 1985, 1989, 2001)[2].

It is envisaged that this meta-evaluation will not only offer valuable insights into a particular evaluation practice - that is, SH's evaluation framework or 'model' - but that it will also offer a means of interrogating the quality or rigour of the results generated via the application of this model. It is hoped that this study might reveal the key strengths - or areas for possible improvement - of the consultancy's model; as well as the quality and effectiveness of the different methodologies that it employs as it seeks to map and document development intervention outcomes and 'change'.

The scope of this study is outlined below:

A. A preliminary investigation of the evaluation *model* itself will be undertaken, focusing upon its components, its relevant theoretical framework and the implications of such a framework / model in terms of its methodological approach.
B. This will be followed by the selection, operationalisation and application of a set of appropriate *quality criteria* by means of a meta-evaluation process – with specific reference to the five outcome assessments / impact evaluations noted above. Here, the focus will be upon interrogating the rigour of the evaluation process and of the research generated via the application of the consultancy's evaluation model.

1.3 Rationale

My selection of this research project is based upon my involvement in the development sector as a social development practitioner. I am employed by SH as a consultant and have thus worked on a variety of the consultancy's evaluations over the course of the past three years, including three of those selected for the purposes of this meta-evaluation[3]. I

[1] The evaluation study procedure utilised by SH includes a series of activities or steps, which is loosely referred to by the consultancy's members as their evaluation 'model'. However, it could also be defined as a framework within which the consultancy conducts all of its evaluation studies and which – to an extent – has become associated with the consultancy itself. For the purposes of this project, the term 'model' will be used to describe SH's evaluation approach and attendant procedures.

[2] Guba and Lincoln were deemed to be especially appropriate to the aims of this research project, owing to a) their naturalistic paradigmatic approach and b) their extensive work within the field of naturalistic evaluation.

[3] The three evaluations in which I was personally involved are as follows: the Outcome Assessment of the LEGO Care for Education Project: "Developing Talents through Creative Play" in Atteridgeville Township; the Evaluation of the Ponahalo De Beers Trust (PDT) Programme and the Evaluation of the United Nations Joint

therefore have a great deal of insight into - and practical experience of - the consultancy's model for evaluation studies. My specific interest in impact evaluations and the various methodologies that may be employed in the measurement and assessment of development intervention outcomes and impact, coupled with the on-going paradigm-based debate regarding the quality and rigour of such evaluation research outputs, led to my interest in undertaking a study of this nature.

Whilst my employment at SH offers me a unique and insightful perspective - plus considerable practical experience - of the organisation's evaluation model, it also presents the risk of possible bias in this research process. This I propose to overcome via the careful selection, conceptualisation and operationalisation of my meta-evaluation criteria. I aim to accomplish this by drawing from a wide variety of evaluation and meta-evaluation sources plus key criteriology texts, to ensure that my quality criteria are both clear and precise, enabling a methodical and standardised application thereof.

I would also argue that a thorough interrogation of the SH evaluation model is of more value to the consultancy than a favourably biased assessment of their work. In keeping with their philosophy of operating as a learning organisation, the consultancy's members encourage on-going critical reflection and debate regarding their work so as to expand and improve upon their skills as evaluators. It is proposed that this research project be approached in a similar manner.

1.4 The research question

The central research question to be addressed by means of this study is therefore as follows:

Can the outcomes and impact of development interventions be assessed with an appropriate level of rigour by making use of SH's naturalistic evaluation model?

As previously mentioned, this question will be addressed via a critical review or meta-evaluation of five SH evaluation studies, as noted in Section 1.2, utilising Lincoln and Guba's (1989, 2001) *trustworthiness* criteria; which are credibility, transferability, dependability, and confirmability.

This set of criteria is proposed by these authors as a means of offering an alternative and more relevant assessment of the rigour and quality of research conducted within a naturalistic / interpretivist paradigm. As such, these criteria are intended to parallel the 'traditional' hallmarks associated with rigorous research; namely internal validity, external validity, reliability, and objectivity, which are associated with the positivist and post-positivist paradigms.

Programme on Human Trafficking (UNJPHT). My tasks varied per project but collectively include project document review, fieldwork, data coding and analysis, and assistance with report writing.

1.5 Methodology

The method selected for undertaking this critical review of SH's evaluation model is the meta-evaluation, which essentially means the 'evaluation of evaluation' (Scriven 2007). The use of a meta-evaluation approach means that the impact evaluations and outcome assessments undertaken by SH will be subject to a standardised review process by means of the application of a predetermined set of criteria to each of the evaluations under review. It is proposed that the utilisation of such an analytical framework will allow for meta-evaluation consistency as well as facilitate a higher level of inter-project comparison and analysis whilst minimising the risk of bias.

As noted by Stufflebeam (1974 / 2011: 135), the meta-evaluation "…should provide retroactive information to help evaluators be accountable for their past evaluation work." However, this research project proposes a slightly different purpose for the meta-evaluation study outlined here. By assessing the quality and rigour of evaluations conducted within the naturalistic / interpretivist paradigm via the application of a set of quality criteria, it is hoped that this meta-evaluation will assist with reflection on - and where necessary improvement of - a specific evaluation model, as well as the development of good practice in terms of the planning, implementing and reporting upon outcome and impact evaluations within the field of socio-economic development. It is also hoped that this meta-evaluation will assist in developing a deeper understanding of the conceptualisation and application of quality criteria for the assessment of naturalistic evaluation practice.

Lastly, it is noted that the meta-evaluation process adopted for the purposes of this project includes a review of secondary data sources; that is, relevant evaluation project documents; plus primary data collection via telephonic, Skype and face-to-face, in-depth interviews. These interviews were conducted with seven members of client organisation personnel, as noted in Table 1, plus the four SH partners.

Chapter 2 Debates regarding evaluation

2.1 Defining and justifying evaluation

Hedler and Gibram (2009) state that *evaluation* may be viewed as both an informal - as well as a formal - undertaking. When conducted on an informal basis, it involves making a value judgement about something, whereas on a more formal basis the term denotes making such a value judgement regarding "…services or professional activities…" (2009: 211). House (1980: 18) concurs with his definition of evaluation as leading to "…a judgement about the worth of something…"; whereas Sanders and Nafziger (2011: 45) note that "Evaluation gives information about the quality of … programs". Morra Imas and Rist (2009: 1) echo the Development Assistance Committee (DAC) Network on Development Evaluation in their definition of evaluation as "…the process of determining the worth or significance of an activity, policy or programme…". Thus it might be argued that within the field of development, the concept 'evaluation' involves a systematic assessment of the worth or quality of a specific intervention, programme, policy or project. In doing so, evaluation is essentially comparative - often explicitly so - as one can only evaluate if one has a set of pre-defined criteria or standards to evaluate against (House 1980)[4].

The discipline of programme evaluation only began in the 1960s and 1970s, particularly in the United States of America, due to the introduction - and rapid expansion - of large-scale social welfare programmes, developed as part of the War on Poverty policy framework (Donaldson and Lipsey 2006; House 1980; Mertens 2009; Posavac and Carey 2007). In seeking to make programme planning more effective and accountable, the United States (US) Federal Government introduced a variety of social science research methods in an effort to enhance the monitoring and quality assessment of its programmes and their services. Now referred to as the 'Golden Age' for research on evaluation, this period witnessed a number of developments in the field, including a focus upon those factors that facilitated the use of evaluations, such as Patton's Utilization-Focused Evaluation (1978). Subsequent to this, much of the work on evaluation was based upon the formulation of models and evaluation theories that prescribe how evaluation should be done.

Donaldson (2009) and Scriven (2003) argue that a second 'boom' in evaluation is currently taking place, but note that it differs to the first in that it is a) more global, b) it is being conducted on a wider range of programmes, practices and organisations, and c) it now includes a higher level of involvement amongst non-governmental and civil society sectors.

So, why should evaluation take place? Patton (cited in Holden and Zimmerman 2009: 16) identifies three uses of evaluation findings, namely the "…rendering of summative

[4] I agree with this assertion and, hence, this is a key consideration in my operationalisation of the meta-evaluation criteria for this project; namely, does the SH model include a clear set of criteria for each of the evaluation studies conducted by the consultancy.

judgements, improving programs formatively, and generating knowledge about generic patterns of effectiveness…". Mark (2008) and Henry (2009) argue that four main evaluation purposes now prevail; namely, programme and organisational improvement, oversight and compliance, assessment of merit and worth, and knowledge development. However, Donaldson and Christie (2005), Donaldson and Lipsey (2006), Bamberger and White (2007), and White (2010) argue that the main requirement of contemporary programme evaluation practice within the development community is *that of determining programme impact*. Such evaluations are generally concerned with the *results* generated by a project, programme or policy as opposed to *process* evaluations, which focus upon the use of inputs, the assessment of activities and outputs against predefined targets, and how well - or efficiently and effectively - a programme is operating (Bamberger and White 2007; Morra Imas and Rist 2009). This shift in focus to impact evaluations and outcomes assessments is due to an increasing emphasis on the achievement of results as embodied in the Millennium Development Goals (MDGs) (Bamberger and White 2007; White 2010). Ton (2012) and Saunders (2011) also note the increasing public pressure on officials and organisations to prove that their use of funding is contributing, in some manner, towards poverty alleviation and improved levels of socio-economic development. Collectively, these influences have given rise to the popular, contemporary 'results-based' evaluation approach.

2.2 Results-based evaluation and the 'paradigm wars'

Whilst results-based evaluation has received widespread endorsement and support in the development sector; it is to some extent being overshadowed by the complexity and number of debates within the evaluation field regarding the most appropriate and rigorous means of defining and measuring intervention results - particularly impact – so as to enable information-generation that is both accurate and useful.

According to the DAC's Summary of Key Norms and Standards (2010: 30), 'impact' may be defined as "The positive and negative changes produced by a development intervention, directly or indirectly, intended or unintended…resulting from the activity; on the local social, economic, environmental and other development indicators." The DAC Network on Development Evaluation notes that such evaluations should also take cognisance of the possible positive or negative impacts of external factors, including fluctuations in economic and trade conditions. Alternatively, 'impact' might be defined as the intended and unintended, long-term[5] change that takes place in organisations, communities or systems as a result of programme activities (Unrau et al 2007: 72); or it might simply refer to the "…final level of the causal chain…" within a programme logic model or programme Theory of Change (White 2010: 154). Eggers (2009: 121) argues that impact is the "…contribution of a project or programme towards the realisation of a higher-level policy goal…"; while the World Bank and International Initiative for Impact Evaluation (3ie) definition of impact is

[5] Unrau et al (2007: 72) argue that the timeframe for impact is 7 to 10 years.

cited as "…the difference in the indicator of interest (Y), with the intervention (Y1) and without the intervention (Y0)." (White 2010: 154) That is, impact = Y1 – Y0 and focuses upon the issue of attribution, or the attribution of observed changes to the intervention under study[6].

The shift to results-based evaluation plus a renewed focus on accountability in terms of the use of aid and donor funding has led towards the re-emergence of the so-called 'paradigm wars'[7] between quantitative and qualitative methodologists, precipitated by the US Department of Education's Institute for Educational Science funding policy prioritisation of experimental – and some types of quasi-experimental – evaluation methods (Coryn 2007; Donaldson and Christie 2005; Henry 2009; and Reichardt 2011). This move caused a split within the American Evaluation Association and was strongly criticised by the European Evaluation Society whilst simultaneously endorsed by the Poverty Action Lab of the Massachusetts Institute of Technology (MIT) and the Centre for Global Development.

Despite the controversy, the 'gold standard'[8] of the randomised control trial (RCT)[9] has re-emerged as the "…ideal embodiment of scientific inquiry…" (Scriven cited in Coryn 2008: 28) and the impact evaluation methodology to which all development organisations and evaluators should aspire; while Ryan-Nicholls and Will (2009: 71) note that the emphasis on results- and evidence-based research is "…again relegating qualitative work to second-class status." Experimental and certain quasi-experimental methodologies are now being posited as a means of ensuring that one's impact evaluations are indeed 'rigorous' and 'objective' (Posavac and Carey: 2007; Picciotto 2012). Picciotto (2012) asserts that the 'popularity of experimentalism' within the development sector is due to the methods' promise of scientific certainty during a time increasingly characterised by challenges, uncertainty and dwindling funding. However, the RCTs "…inherent limitations in applied settings and the potential for misuse have become evident." (Posavac and Carey 2007: 25) These limitations include the following:

[6] A number of evaluation theorists argue for the use of contribution analysis as opposed to trying to determine the attribution of a particular intervention to particular observed outcomes or changes. As noted in Patton (2008: 4), contribution analysis focuses upon identifying the causes or level of influence that an intervention is likely to have had, as opposed to simply addressing the issue of 'Did X cause Y?'

[7] The current debate has been dubbed the 'causal wars' by Scriven (2008) as it centres around the key question of what counts as scientifically 'impeccable evidence' of a causal connection between intervention and change.

[8] In medicine, a 'gold standard' test refers to a diagnostic test or benchmark that is regarded as definitive.

[9] The randomised control trial is defined as the use of an experimental design, involving at least two groups of subjects; namely, the control group and the experimental group, which may also be referred to as the study or treatment group; where these subjects are distributed by a strictly random process and which are not further identified / distinguished by any common factor besides the application of treatment to the experimental group (Scriven 2008: 11).

a) Development projects are often large, complex and heterogeneous[10], whilst many evolve and adapt over the course of implementation. Furthermore, different implementing agencies may be involved and hence it is extremely difficult to ensure any level of uniformity in terms of how the project is implemented.
b) RCTs can seldom take into account the setting in which the programme is being implemented and it is often not possible to control the economic, political and social milieu sufficiently – and over the entire course of the long-term - to allow for a 'controlled implementation setting'.
c) Experimental methods do not allow for the study of how the implementation process may have affected the outcomes; and lack flexibility in terms of being able to identify, study and report upon changes that may have taken place – both in project administration and in the project setting – during the life cycle of the project.
d) There is the risk of 'seepage' or 'contagion' of the control group, who may become part of the project population due to social interaction or proximity to the intervention or programme sites.
e) Professional experience indicates that RCTs are more expensive to run, plus they are often deemed to be inappropriate due to ethical concerns, such as the deliberate withholding of medical treatment from control group members who may urgently require such treatment. As noted by Bickman and Reich (2009: 62), "…if it is known with some degree of certainty that one treatment is better than another, then one must question why the study is to be conducted."

Nevertheless, Cook (2006), Bamberger and White (2007), and White (2010) argue that there are a number of situations where RCTs can – and should – be applied. This includes evaluations where there is a clearly defined target group; where it is possible for random allocation to a treatment and a control group to take place; where the 'treatment' can be applied in a uniform and standardised way; and in instances where the project setting remains fairly stable throughout the period of intervention. Thus RCTs are deemed to be best suited to trials / research studies that are relatively "…simple, bounded and linear…" (Patton 2008: 4) and which are conducted over the short term. Despite the narrow scope for such research studies in the complex arena of socio-economic development, Bamberger and White (2007: 62) argue, that there are still too few development projects that are being "…subjected to rigorous impact evaluations, and the vast majority have been assessed without even a simple quasi-experimental design or any reference to a counterfactual."

Based upon the above, I concur with Mertens' (2009) and Christie and Fleischer's (2009) assertion that two main 'camps' or paradigms have emerged in evaluation research. Although these authors' dichotomous classification and description of these 'camps' might

[10] Rogers (2008) offers an interesting categorisation of interventions as being 'simple', 'complicated' and 'complex'. For further input, please refer to the text noted under 'Sources Consulted'.

be accused of over-simplification[11], it offers a sound basis upon which this research project may rest. The table below offers a summary of the two main evaluation research 'camps', as proposed by the abovementioned authors; that is, the *post-positivist* and the *constructivist* paradigms.

Table 2: The post-positivist and constructivist paradigms

Post-positivist	**Constructivist**
Assumption is that there is a single reality that exists and that can be studied objectively; whilst recognising that a 'full' understanding of such a reality or 'truth' is not possible. A rigorous study or one of good quality is that which most accurately reports upon – and re-presents – such a reality.	Assumption is that multiple explanations of reality exist which are subjective and hence can change according to the 'knower'; that is, realities are constructed based upon the individual's past experiences and current contexts.
Deductive reasoning and quantitative research methods underpin post-positivist research strategies.	Inductive reasoning and qualitative research methods underpin constructivist research strategies.
Cause and effect are linked; whilst causation is observable and measurable, but with a level of uncertainty. Seeks to attribute observed change in outcome indicators to a specific intervention.	Cause and effect are believed to be impossible to distinguish as relationships tend to be more complex and bidirectional, with multiple influences being experienced simultaneously.
Use of theory to establish a priori hypothesis against which a conclusion will be drawn. Generalisability from such a conclusion is the aim of the research.	Particular instances are studied and outcomes from such research are used to infer broader or more general laws. However, generalisability is not perceived as being as important as local relevance.

Lately, a third 'camp' has been established within the evaluation sector. Patton (2002), Christie and Fleischer (2009), and Saunders (2011) argue for a third paradigm for evaluation and applied research, namely that of *methodological pragmatism.* These authors propose pragmatism as a possible compromise between the two paradigms noted above in that it

[11] To simply note a post-positivist versus a constructivist paradigm is to overlook the myriad paradigms and approaches, plus methodologies, which exist in contemporary research – specifically within the broad categories of 'evaluation research' and 'qualitative research'. However, such a distinction serves the aims and objectives of this research project.

acknowledges both quantitative and qualitative methods as legitimate forms of enquiry[12], whilst supporting the concurrent use of both deductive and inductive logic. Adopting such a pragmatic approach implies that the selection of an appropriate research method will not be based upon the researcher's or evaluator's paradigmatic preference, but rather upon the nature of the study in question[13]. As stated by Patton (2002: 72), being pragmatic "…allows one to forego methodological orthodoxy in favour of methodological appropriateness…". In a similar vein, the judgement or assessment of the quality of such a study will not be made on the basis of its use of a counterfactual, but rather according to its methodological 'responsiveness' to the intended aims and objectives of the study, and the resources available for such a purpose. Hence, as a means of reconciling the two opposing methodological 'camps' noted above, the use of a pragmatic approach incorporating the use of mixed research methods[14] (where appropriate) for outcomes assessments and impact evaluations has been proposed. Picciotto (2012) refers to this 'uneasy truce' as showing great promise, but notes that it is still very much in its infancy.

An additional factor to be considered in the burgeoning break from methodological orthodoxy is that current evaluation practice includes the use of more participatory evaluation methods and, as a result, the selection and use of an appropriate study methodology will generally be made in consultation with the relevant stakeholders – and not by the evaluation team alone (Donaldson and Lipsey 2006). Based upon my experience in the development sector, it appears as though the methodologically pragmatic approach is certainly one currently favoured amongst client organisations – as well as by SH itself.

2.3 Assessing the rigour of evaluation research

The concept 'rigour' is variously defined and conceptualised; for example, Reynolds (1971, quoted in Ryan-Nicholls and Will, 2009: 70) defines rigour as "…the use of logical systems that are shared and accepted by relevant scientists to ensure agreement on the predictions and explanations of the theory…", while the Glossary of Key Terms of Colorado State University defines the term as the "Degree to which research methods are scrupulously and meticulously carried out in order to recognise important influences occurring in an experiment." Less positivist and more in keeping with a methodologically pragmatic approach is Ryan's (2009: webpage, no page number indicated) definition of rigorous research as being "…research that applies the appropriate tools to meet the stated objectives of the investigation and which is both transparent and explicit…"; while Krefting

[12] As opposed to the current 'methodological hierarchy', where either quantitative or qualitative research methods are perceived as one being superior to the other.

[13] Seale (1999) also advocates a research practice that can operate autonomously from a simple adherence to a specific philosophical or theoretical perspective. This adherence, he argues, sets in place a number of constraints as the researcher is bound to make methodological decisions which exist only within the parameters of his / her selected position.

[14] A mixed methodology involves the use of both qualitative and quantitative data collection and analysis methods - either simultaneously or sequentially - as part of the same study.

(1991) speaks of rigour as simply referring to the 'worth' of a research project. As noted by Coryn (2007), a high number of the members of the scientific community associate the concept 'rigour' with research quality, credibility and accuracy. Therefore, accusing a research project of a lack of rigour implies that it has delivered results that are untrustworthy and poor in quality.

In terms of assessing the rigour of evaluation research, the so-called 'Holy Trinity' of quantitative research; namely validity, reliability and objectivity, generally applies to research endeavours undertaken by those working within the positivist / post-positivist paradigm (Coryn 2007). However, this level of clarity and consensus regarding hallmarks of rigorous practice appears to be lacking within the naturalistic / constructivist research camp. Here, a long-standing and extensive 'criteriology' debate exists regarding the feasibility and desirability of establishing quality criteria for the assessment of constructivist research, which is predominantly undertaken via qualitative methods (Sandelowski 1993, Seale 1999, Spencer et al 2003, Tracy 2010). Positions in this debate range from a total rejection of the notion of quality criteria for qualitative research[15] (Smith 1984, 1990 cited in Spencer et al) to those who advocate the use of criteria common to both naturalistic and positivist inquiry (Morse et al 2002, Le Compte and Goetz 1982, Kirk and Miller 1986). Mid-way between these two extremes lie those who do not reject the notion of quality criteria but who propose that criteria formulated for the purpose of conducting quality assessments should be based upon the ontological and epistemological assumptions underpinning each philosophical and methodological approach.

Occupying such a 'mid-way mark' in the criteriology debate are Guba and Lincoln (1981, 1985, 1989, 2001) who do not reject the notion of quality criteria nor their application as hallmarks of rigorous research practice. Instead, these authors propose four factors that relate to tests of rigour across all forms of research - specifically truth value, applicability, consistency, and neutrality - but argue that these four fundamental factors should be interpreted, conceptualised and applied according to the philosophical framework within which the researcher is working. Guba and Lincoln (1985, 1989) therefore assert that the so-called 'traditional' or positivist criteria for assessment of rigorous research practice – that is, validity, reliability and objectivity – are not appropriate for those operating within a more naturalistic and constructivist paradigm. As such, they proposed a set of 'trustworthiness' criteria to conceptually 'parallel' the rigour criteria of positivist inquiry, but which "…allow for the changed requirements posed by substituting constructivist for positivist ontology

[15] Those who reject the notion of quality criteria for constructivist / qualitative research do so for a number of reasons. These include the anti-foundational and relativist premises upon which many qualitative approaches are based; namely, that there are many different constructions of reality, so no 'correct' or singular depiction thereof can exist (Sandelowski 1993). Other objections are based upon the perception that the use of any form of criteria and assessment will inhibit the creativity and flexibility of qualitative research, which is deemed to be a key strength of the approach (Ryan-Nicholls and Will, 2009); while Rolfe (2006) and Dixon-Woods et al (2004) argue that any attempt to formulate a generic set of quality criteria for qualitative research is not feasible due to the wide range of theories and methods included within such research.

and epistemology…" (Guba and Lincoln 1989: 236). This, they argue, will ensure that qualitative research is judged 'on its own terms' or according to standards or criteria more suited to its ontological and epistemological 'position'[16]. The table below offers a summary of this discussion.

Table 3: Hallmarks or criteria for assessing rigorous research practice

Aspect	Positivist criteria	Lincoln and Guba's 'parallel' trustworthiness criteria
Truth value	Internal validity	Credibility
Applicability	External validity or generalisability	Transferability
Consistency	Reliability	Dependability
Neutrality	Objectivity	Confirmability

Situating myself within the criteriology debate, I concur with Spencer et al (2003: 70) who argue that not only is the assessment of evaluation research quality and rigour necessary, but that "…all aspects of rigour are of *greater importance* in evaluative studies because of their explicit purpose of informing policy-making, and thus contributing to change which will have real impacts on people's lives." I propose that the setting of standards for good evaluation practice, via the use of quality criteria, will encourage methodological reflection, contribute towards on-going improvement in the quality of evaluation work being conducted, plus enhance the legitimacy of one's work within the development and evaluation sectors. Furthermore, I propose that the absence of quality criteria for qualitative evaluation research may lead to the judgement thereof according to quantitative hallmarks of rigorous practice - specifically against the 'gold standard' of RCTs as previously discussed. Such a judgement would inevitably deem qualitative work to be of substandard quality and value.

[16] Spencer et al (2003) note that Lincoln and Guba's trustworthiness criteria are often considered as the 'gold standard' for quality assessments of qualitative research. However, it should be noted that their work has been contested and fiercely debated, and has provoked a considerable amount of criticism from both quantitative and qualitative methodologists. Lincoln and Guba responded to such criticism by introducing a fifth criterion of 'authenticity', which was deemed to be more responsive to the concerns of constructivists - unlike the four trustworthiness criteria, which are methodology-focussed and which arose in response to positivist conceptualisations of research quality and rigour (Pike Hall 1995, Guba and Lincoln 1989). Given the focus of this research project, it was elected to focus upon those four criteria that spoke specifically to methodology and the assessment of rigorous practice. For this reason, the fifth criterion of 'authenticity' will be excluded from the meta-evaluation framework for this research project.

Thus, I support writers such as Guba and Lincoln (1985, 1989) who argue that if quality and rigour assessments are conducted, this must be done via the application of relevant and appropriate criteria. In light of this discussion - and given that the SH evaluation model falls within the naturalistic / constructivist paradigm and strongly supports the use of qualitative research methods - I propose to assess the rigour and quality of the consultancy's evaluation process and outputs via the application of a set of criteria based upon the writings of Guba and Lincoln (1985, 1989). This argument will be justified by the description of the SH evaluation model in the chapter that follows.

Chapter 3 Southern Hemisphere's evaluation model

3.1 Southern Hemisphere – the consultancy

Southern Hemisphere (SH), a Cape Town-based socio-economic development consultancy, was established in 1999. As noted on the SH website, the consultancy specialises in "…participatory development interventions with a focus on learning organisations…"[17]; and assists a wide range of national and regional[18] state, non-state and donor institutions with applied research; programme design, monitoring and evaluation; capacity building in the form of public and in-house training and facilitation; and organisational development.

Following a number of organisational changes, the consultancy currently consists of a core team of four partners as well as number of consultants and research associates with whom they collaborate on an on-going basis.

3.2 Southern Hemisphere – the evaluation model

Based upon the interviews conducted with the four SH partners, as well as my personal experience in working for the consultancy, I would argue that the SH model for undertaking outcome assessments and impact evaluations might be described as a naturalistic-pragmatic model:- *naturalistic*, in that the consultancy focuses upon grounded inquiry or study within the 'real-world' as opposed to experimental or manipulated settings, whilst acknowledging that such a reality is comprised of multiple perspectives, which are both divergent and inter-related; and *pragmatic* in that the evaluation aims and objectives – plus resources available for the study – will influence methodological decisions. However, the model is described by all of the consultancy's partners as being first and foremost a *participatory process*, underpinned by broad consultation and engagement with a variety of stakeholders affiliated to the programme or project under review. This approach is based upon three key considerations, endorsed by all four of the consultancy's partners; namely:

a) The ethical imperative of the evaluator to ensure that all relevant parties are offered the opportunity to give their input; that is, to ensure equality of representation amongst all relevant stakeholders, including programme management, programme implementers, target groups, and beneficiaries. This is highlighted in the quote below:

   "The SH model emphasises the value of the perspectives of the beneficiaries themselves. For this reason, we strongly advocate that they are included in the evaluation and that their perspectives and opinions are heard. I would say that the beneficiaries' input

[17] Source: www.southernhemisphere.co.za. Accessed 12/12/2012 and 14/04/2013.
[18] The consultancy's work has been predominantly national; however, it is becoming increasingly involved in regional / Southern African projects; including countries such as Namibia and Zambia.

almost carries more weight than that of the programme implementers. They are central to the way that we undertake our evaluations." (SH partner 4)

b) Maximum involvement of the client organisation to ensure that its evaluation requirements, constraints and information needs, are clarified and addressed; and

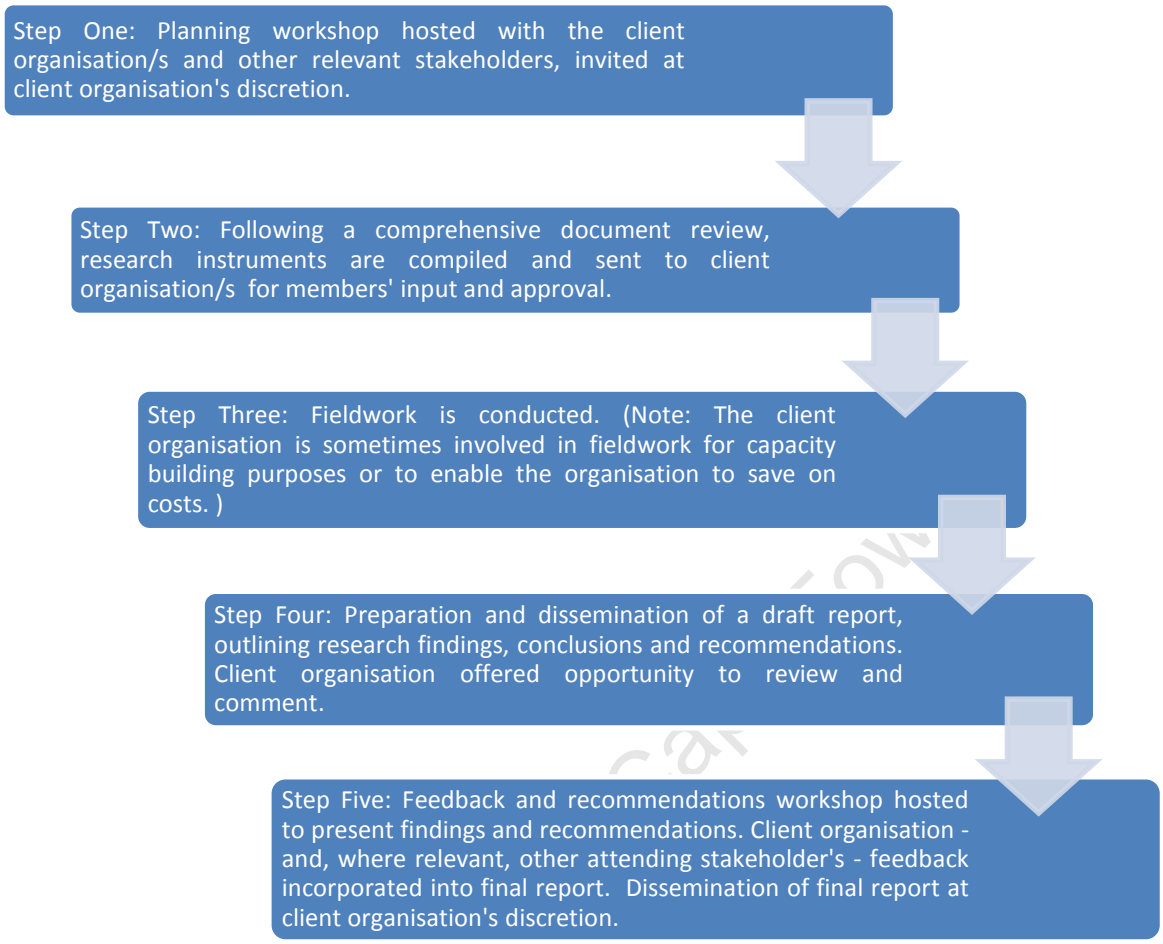
c) Empowerment of the client organisation, which is offered insight into evaluation processes, methodologies and benefits, as well as the skills to act upon such knowledge in future evaluations. This is in keeping with the consultancy's belief that programme evaluations offer a valuable means of facilitating reflection, learning and organisational development and, as a result, the client organisation is encouraged to view the evaluation process not as an accountability measure only, but as a means of improving upon its work and effectiveness[19]. As noted by Stufflebeam (1974), 'accountability' is a highly threatening concept for most development organisations. To counteract the stressful nature of such evaluations, and possible subversion amongst those being evaluated, Stufflebeam (1974) advocates the inclusion of client organisation stakeholders in all aspects of the evaluation, such as the design and planning thereof. This clearly articulates with SH's approach of client engagement and empowerment.

To date, this evaluation model has not been explicitly defined in any way. The general perception amongst the partners is that the model is not rigid or prescriptive; but that it operates as a 'framework' to inform and guide their evaluation processes. This allows for flexibility for both the consultancy as well as for the client organisation. Nevertheless, underpinning the model is a number of established procedures. This includes the flexible use of the DAC standards as a set of criteria against which the intervention's strengths and challenges may be recorded and assessed[20]. Secondly, the process is clearly divided into a number of stages or activities, which collectively form a coherent, linear process or 'mode of operation'. These are outlined in the graphic below and are elaborated upon in the section that follows.

[19] Such an approach is referred to by Bawden and Packham (cited in Patton 2002: 179) as 'systemic praxis'.

[20] These criteria are as follows: *Relevance*, *Sustainability*, *Effectiveness*, *Efficiency,* and *Impact*. Chianca (2008: 41) argues that these evaluation standards have been "…by far the most influential work in the field of development evaluation…" because they shifted the focus away from outputs and project economic rate of return towards a broader set of key elements, including intervention outcomes and impact. Thus it might be argued that the use of the DAC criteria as a basis for their evaluation model indicates that SH is functioning well-within contemporary results-based evaluation discourse.

Diagram 1: An outline of the Southern Hemisphere Evaluation Model



The planning workshop, as noted by all of the SH partners, is a pivotal part of the evaluation process as it is at this particular stage that the evaluation aims and scope, methodology, instruments, and timeframes are clarified and confirmed[21]. In addition to this, discussions are held during the planning phase regarding the programme / intervention's logic and aims, thus offering the consultancy an appropriate level of background knowledge and insight into the project under review[22], which in turn greatly benefits the overall planning, implementation and outcomes of the evaluation process. This is facilitated by a concurrent

[21] It was noted that recent practice included SH's compilation and dissemination of an Inception Report, including all planning workshop details. This was recognised as an important means of confirming the way forward in the evaluation process, so as to prevent miscommunication and possible client dissatisfaction.

[22] In certain instances, the client organisation is also assisted with the formulation of a Terms of Reference for the evaluation or with the formulation of a programme / intervention logic model. The design of such a logic model and its attendant objectives-based evaluation, is an approach which "…emphasises the use of clearly stated programme goals and objectives…", (Posavac and Carey 2007: 26), as a means of measuring programme achievements. This approach is often criticised due to the possible neglect of other relevant issues, such as other influencing variables or unintended and possibly negative outcomes. However, SH maintain that they avoid such oversights via their adoption of a flexible approach and their inclusion of as many different stakeholders as possible in their evaluation process.

or subsequent comprehensive document review of all relevant organisation / programme material.

The selection of an evaluation methodology is guided by the purpose and aims of the evaluation, the preferences and information needs of the stakeholders, the timeframe within which the evaluation needs to be completed, and available resources. Posavac and Carey (2007) strongly endorse such an approach, arguing that the selection of an appropriate methodology or study design must address the needs of the stakeholders and not be the routine or preferred choice of the evaluator.

Following the planning workshop and document review, the evaluation research instruments are compiled and sent to the client organisation for its members' input and approval. Thereafter, fieldwork is coordinated and conducted. It was noted that members of the client organisation were periodically involved in data collection and fieldwork for capacity building purposes or to allow the client organisation to save on costs[23]. Thereafter, data analysis is conducted followed by the preparation of a draft report, outlining the research findings, conclusions and recommendations. This report is disseminated to as many stakeholders as possible. Once again, the client organisation/s is offered the opportunity to review and comment upon the material submitted. To facilitate this process, a feedback and recommendations workshop is hosted by SH where evaluation findings are presented and recommendations are formulated through dialogue with the client organisation, together with any other relevant stakeholders[24].

The use of such a feedback and recommendations workshop is noted by Miles and Huberman (1994: 275) as one of the "…most logical sources of corroboration…", on the basis that the actor within the setting under study is "…bound to know more than the researcher ever will about the realities under investigation." Guba (1981) also endorses such an approach as a means of undertaking a so-called 'member check'. Following the feedback and recommendations workshop, SH prepares a final evaluation report for delivery to the client.

This particular process or 'model' is followed for all impact evaluations and outcome assessments conducted by the consultancy. Whilst some flexibility is afforded to client organisations on the basis of possible time or budget constraints, the basic steps remain essentially the same. The reason for this is explained below[25]:

[23] However, it was also noted in the course of the interviews with the SH partners that this was undertaken infrequently due to the possible risk of bias in the respondent input obtained. Alternatively, such involvement by the client organisation was restricted to data collection for baseline study purposes only.

[24] Participation at the feedback and recommendations workshop is at the discretion of the client organisation, and is generally influenced by time and budget constraints. As such, the SH partners noted that unfortunately feedback to – and inclusion of - beneficiary groups was, in their experience, generally limited. However, it was also pointed out that SH might offer suggestions and input regarding those selected and invited to attend.

[25] All of the SH partners noted that divergence from their basic model was not without risk and that the absence of key processes could impact negatively upon the evaluation itself.

> *"Without this process, things often go wrong – I have learnt this along the way. For example, with PDT we had another consultant conduct the planning meeting in Johannesburg on our behalf. As a result, I felt that I lacked important information and insight into the project; for example, I was initially not aware of their use of implementing agents." (SH partner 1)*

The evaluation model outlined above is also described by the consultancy's partners as an 'emergent' one, given that – following its initial conceptualisation by the founding member of SH - it has been continuously adapted and refined over the course of the consultancy's history. The partners noted that the model is not based upon any specific theory of development or applied evaluation research, but rather upon practice, experience and the reflexive dialogues of the consultancy's members in conjunction with the associates and clients with whom they work. Other influences cited include collaboration with academic institutions and other development agencies. This is in keeping with the findings of a study conducted in 2003 by Christie (as outlined by Dillman 2013), which indicates that less than 10 per cent of the evaluators sampled for the study make use a particular theory to guide their work.

Despite the fact that the model was not specifically theory-driven, all of the partners felt that appropriate theoretical bases[26] were inherent within the model, but noted that these had generally been discovered retrospectively. The following quotes elaborate on this:

> *"I wouldn't say that it is based on a specific theory really. Practice has emerged from practice and not from theory. Perhaps I would say that it is a utilisation-focused approach located around participation and building learning in the client organisation – how will this help to inform their practice, etcetera. This is especially relevant when it comes to the feedback and recommendations workshop. It is also about learning for us." (SH partner 1)*

> *"It's not like we read up about evaluation theory and then said 'Oh, this is a theory that we like and so this is our model'; but we have come across theoretical texts in the course of our work that we have discovered do actually underpin and 'speak to' or explain what we do and so this is where we have managed to locate or situate our model." (SH partner 2)*

Initially, SH focused upon the use of qualitative research methods. However, it was noted in the course of the interviews conducted for this research project that the consultancy has recently experienced a move towards the increased utilisation of mixed research methods[27]. However, this has not impacted negatively upon the consultancy's use of qualitative

[26] Theories mentioned in the course of the interview process with the four SH partners included asset-based community development, participatory inquiry, utilisation-focussed, and empowerment evaluation.

[27] This move was partly due to an increased interest in such methods amongst the consultancy's client base, coupled with the partners' awareness of the growing shift in the development sector towards the incorporation of such methods in programme evaluations.

research methodologies; which were deemed - by all of the SH partners - to be a fundamental part - and key strength - of their evaluation model.

3.3 Evaluation model assumptions

The assumptions underpinning the SH evaluation model may be summarised as follows:

Diagram 2: A summary of the assumptions underpinning the SH evaluation model

Chapter 4 Conceptualisation and operationalisation

4.1 The meta-evaluation

With evaluation now a "…mandated activity…" (House 1980: 15), there has been an increasing level of interest in evaluating evaluation itself. As stated by Stufflebeam (1974, republished 2011: 99), "…good evaluation requires that evaluation efforts themselves be evaluated." However, Dillman (2013) notes that the evaluation field places evaluators under a great deal of pressure to *do* evaluation – not to *study* it (my italics). This has left the field of meta-evaluation relatively unattended to for the past four decades.

Nevertheless, the past few years have witnessed an increasing body of work and study located around meta-evaluation practices and results. It might be argued that this is a positive and noteworthy step forward in terms of establishing appropriate standards, guidelines and criteria for improving evaluation practice within the development sector – as opposed to the on-going causal wars and methodology supremacy battles as previously discussed.

Scriven was the first to use the term 'meta-evaluation' in 1969 to refer to the evaluation of evaluation (Sanders and Nafziger 2011). Cooksy and Caracelli (2009: 2) later defined meta-evaluation as the "…systematic review of evaluations to determine the quality of their processes and findings…", while Stufflebeam (2011: 135) defines meta-evaluation as "…a procedure for describing an evaluation activity and judging it against a set of ideas concerning what constitutes good evaluation." As noted above, meta-evaluations may best be described as a fluid and 'emerging' concept. As such, their use, definition and scope vary - sometimes considerably - from one author to another. However, it would appear that there are some areas upon which most evaluation authors agree, notably that:

a) Meta-evaluations may be conducted internally as well as externally, and for formative and summative purposes.

b) As such, meta-evaluations may be used proactively to improve upon decision-making and provide recommendations; or retroactively for accountability purposes and to enable the assessment of evaluation merit.

c) Meta-evaluations may be used to assess a single study or multiple evaluations; to synthesise evaluation results of a particular intervention; or to assess the evaluation capacity of a particular group or organisation.

Hedler and Gibram (2009: 219) note that "…the meta-evaluation depends upon a set of quality criteria to make it valid." As such, a variety of suggested meta-evaluation criteria or evaluation standards abound. These have been compiled by evaluation associations, such as

the African, American and European Evaluation Associations; well as by leading evaluators and evaluation theorists, including Patton, Scriven, Stufflebeam, Mark, and Henry (Cooksy and Caracelli 2009). While these frameworks informed the operationalisation and assessment processes conducted for the purposes of this project, the four trustworthiness criteria proposed by Guba and Lincoln (1989) formed the foundation of this meta-evaluation.

4.2 Conceptualisation of the meta-evaluation criteria

As noted above, the literature review uncovered a large number and wide variety of criteria 'frameworks' for assessing the quality of qualitative evaluation research and qualitative research in general[28]. Of note is that none of these frameworks are presented as being definitive, exhaustive, comprehensive, or prescriptive. Rather, they are described as *guidelines* for assessment, 'critical appraisal *checklists*', or simply as a 'list of questions' to guide quality appraisals. The rigid and mechanistic specification and application of quality criteria is generally denounced in favour of the provision of a flexible 'tool' for appraisal purposes. This, the majority of the framework compilers argue, is due to the flexible, iterative and multifaceted or varied nature of naturalistic evaluations and qualitative research, which do not lend themselves to a set of generic, formalised indicators against which all such research projects might be measured.

Whilst I concur with this argument, this lack of formalisation and operationalisation of specific quality indicators, coupled with limited evidence of any practical or systematic application of criteria frameworks to evaluation and qualitative studies, made the same task within this research project all the more challenging[29]. Compounding this was the potential for bias created by my close involvement with SH and its partners.

As noted, the work of Lincoln and Guba (1981, 1985, 1989, 2001) was utilised extensively when compiling the meta-evaluation framework for this research project. However, a wide variety of additional sources were consulted for the purpose of formulating clear definitions and distinct quality indicators for each meta-evaluation criterion. It was hoped that this process would allow me to develop a meta-evaluation framework that was itself both 'rigorous' and of good quality.

An overview of the initial step in the development of my framework is offered below; that is, a comparison of various trustworthiness criteria definitions - and proposed tools for the assessment of each criterion. This step was followed by the construction of a final set of definitions and tools where clear overlaps - and sound justifications – for a particular

[28] The following sources were consulted: African Evaluation Association 2005, American Evaluation Association 2005, Bamberger and Rugh 2008, Centres for Disease Control 1999, Joint Committee on Standards for Educational Evaluation 1995, Patton 2003, Sanders and Nafziger 2011, Scriven 2007, Spencer et al 2003, Stufflebeam 2011, Tracy 2010.

[29] Spencer et al (2003) and Dixon-Woods et al (2004) note similar findings.

definition or quality assessment indicator emerged from my reading of the consulted resources.

Table 4: Step 1 in development of the meta-evaluation framework: comparison and assimilation across sources

Criterion	Proposed definitions	Proposed means of assessing - obtained via literature review
Credibility	Credible evidence is generally defined as that which is perceived by stakeholders to be trustworthy, plausible / believable and relevant to the aims and objectives of the evaluation. Tracy (2010: 842-843) defines credible findings as those that "…readers feel trustworthy enough to act on and make decisions in line with."	Evidence of *member checks* - also referred to as 'respondent validation' and 'member reflections' - is proposed as the primary means of assessing credibility (Dixon-Woods 2004, Guba and Lincoln 1989, Seale 1999, Tracy 2010). Such a member check involves the "…sharing and dialoguing with participants about the study's findings and providing the opportunity for questions, critiques, feedback, affirmation and collaboration." (Tracy 2010: 844) Other options for assessing the level of credibility of a study include: evidence of the use of purposive sampling to enable *negative case analysis* (Barbour 2001, Guba and Lincoln 1989); the use of *prolonged engagement* at the site of inquiry, *peer debriefing* and *progressive subjectivity* (Guba and Lincoln 1989); and *methodological, investigator and data source triangulation*[30] (Durrheim and Wassenaar 1999, Seale 1999).
Transferability[31]	The applicability or 'fittingness' of the evaluation findings, conclusions and recommendations to other contexts or programmes of a similar nature. Tracy (2010: 845) defines the term as the "…study's potential to be valuable across a variety of settings and situations."	Determining the degree of transferability of the research is primarily the responsibility of the one doing the transferring. As noted by Trochim (2001: 162) "The person who wishes to transfer the results to a different context is then responsible for making the judgement of how sensible the transfer is." Guba and Lincoln (1989), and Durrheim and Wassenaar (1999) assert that the research report must include an extensive, detailed or 'thick' description of the research context to enable the reader to undertake an adequate assessment of the transferability of the research findings and recommendations.

[30] It should be noted that Guba and Lincoln do not advocate the use of triangulation as a means of assessing credibility. This is due to the authors' association of triangulation with the positivist notion of a set of unchanging phenomena, which might serve as the 'fixed point' against which other data might be checked. However, I concur with Seale (1999: 472) who argues that triangulation is a "valuable craft skill" and that it has a place "within a variety of paradigms."

[31] The question of 'transferability' is a contested one, given that most forms of evaluation are concerned with measuring the "…merit, worth and / or significance of a particular program…" (Scriven 2008: 19) and the findings thereof are - generally speaking - not applicable to a broader population, nor are they expected to be. Alkin and Stake, (cited in Donaldson and Lipsey 2006: 63) argue that evaluators should be more concerned about producing local and highly relevant knowledge as opposed to generalizable knowledge. However, this criterion was included for the purpose of assessing the relevance and usefulness of the outputs as most of the client organisations included in this study are operating on a national or even global level, with a number of concurrent and similar programmes / operations in place.

Dependability	Refers to the underlying logic and defensibility of findings, conclusions and recommendations; hence addresses questions regarding the consistency, coherence and stability of the study process and data analysis over time and across researchers and methods.	Guba and Lincoln (1989: 242) note that the concept dependability is concerned with the "…stability of data over time." However, they point out that dependability excludes any changes that may occur because of the evaluator's 'maturing reconstructions', which are expected in a data analysis approach that is iterative, inductive and emergent. The proposed means of assessing dependability is the *dependability audit* which is essentially a judgement of the inquiry process and the interpretations that emerge as a result thereof (Durrheim and Wassenaar 1999, Guba and Lincoln 1989, Seale 1999). Miles and Huberman (1984, 1994) propose that auditability can be achieved by making all research material, including personal memos, data coding sheets and display charts, available for external scrutiny; that is, the researcher must provide rich and detailed descriptions to demonstrate how reported actions and opinions emerged or developed from the collected data. Tracy (2010: 841) notes that rigorous data analysis is "…marked by transparency regarding the process of sorting, choosing and organising the data." Hence, the research report (particularly the methodology section) is suggested as the primary location for auditability.
Confirmability	Findings should be drawn from available data and should be free / independent of the perceptions and possible biases of the evaluators; that is, they should be deemed to be value-neutral and representative of all stakeholders' input. Trochim (2001) and Tracy (2010) assert that the term refers to the degree to which the (evaluation) results can be confirmed or corroborated by others.	Procedures suggested for assessing the level of confirmability include *multiple coding*, similar to quantitative inter-rate reliability, where coding and data analysis are 'checked' by independent researchers (Barbour 2001). However, Barbour asserts that the real potential of multiple coding lies in its provision of alternative interpretations of data and its encouragement of critical reflection amongst those undertaking a study. Guba and Lincoln (1989) propose the use of a *confirmability audit*, which includes a close scrutiny of original data sources, the logic of the analysis of the research data, and how these link with the final conclusions. This overlaps considerably with their proposed dependability audit outlined above.

4.3 Operationalisation of the meta-evaluation criteria

Following an extensive literature review, including a scrutiny of 10 evaluation / meta-evaluation checklists which focused upon naturalistic and / or qualitative inquiry, I formulated a final definition for each of the four meta-evaluation criteria, as well as a set of quality indicators for each criterion. An outline of the finalised meta-evaluation framework is offered below.

Table 5: Finalised meta-evaluation framework, including criterion definitions and quality indicators

Key Criteria	Definition	Quality Indicators
1. Credibility	Evaluation evidence is perceived by stakeholders both internal and external to the study to be trustworthy, plausible, useful, and relevant to the aims and objectives of the evaluation.	1.1 Evidence of member checks / reflections on research design, process, findings and recommendations. 1.2 Evidence of investigator and/or data source and/or methodological triangulation. 1.3 Evidence of external review of the research process and findings. 1.4 Rationale and description of sample composition and selection (for example, characteristics of sample, basis for inclusions and exclusions, sample size and how sample allowed comparative data and / or negative case analyses to be undertaken). 1.5 Client feedback indicates that the SH outcome and / or impact evaluation outputs are perceived as being trustworthy, credible and believable. 1.6 Client feedback indicating that the SH outcome and / or impact evaluation outputs are perceived as being relevant and useful, incorporating supportive evidence; that is, examples of client utilisation of evaluation findings and recommendations.
2. Transferability	The applicability or 'fittingness' and value of evaluation findings, conclusions and recommendations to other contexts or programmes of a similar nature.	2.1 Evidence of 'thick', detailed descriptions of research context to enable readers' assessment of transferability. 2.2 Evidence that client organisations regard SH evaluation outputs to be of value in / transferable to other contexts or programmes of a similar nature.
3. Dependability / Auditability	The consistency, coherence, logic and stability of the study process and data analysis over time and across researchers and methods.	*3.1 Dependability audit:* 3.1.1 Inclusion of detailed audit trail of data collection process and procedures thereof. 3.1.2 Inclusion of detailed audit trail of data analysis process and procedures. 3.2 Documentation of any changes to evaluation design, including reasons and any implications thereof. 3.3 Documentation of any study limitations and their possible impact upon the evaluation. 3.4 Evidence that client organisations regard the SH evaluation outputs to be stable and replicable across researchers and methods.

4.Confirmability	Findings are drawn from available data and are independent of the perceptions and possible biases of the evaluators; that is, they are deemed to be value-neutral and representative of all stakeholders' input.	*4.1 Confirmability audit:* 4.1.1 Evidence of use of multiple fieldworkers /interviewers. 4.1.2 Indication that fieldworker training was conducted. 4.1.3 Evidence of the use of appropriate and unbiased questioning techniques in data collection instruments. 4.1.4 Evidence that data collection instruments were piloted. 4.1.5 Indication that quality checks were conducted on submitted transcripts. 4.1.6 Indication of the use of multiple coders for inter-coder cross-checks. 4.1.7 Use of peer review of coding and data analysis. 4.2 Use of a set of pre-defined and clearly stipulated criteria against which the programme /intervention will be assessed. 4.3 Evidence that client organisations regard the evaluation process and outputs as being sufficiently neutral and inclusive of all stakeholders' input.

Chapter 5 Methodology

5.1 Qualitative methodology

A qualitative methodology, incorporating the use of a meta-evaluation, was selected for this study. As previously noted, a meta-evaluation analytical framework was developed for the purpose of this research, including a clear definition of each of the selected quality criteria plus a set of quality indicators for each criterion, to ensure standardisation of the meta-evaluation. The framework also operated as the basis for my instrument design, data coding and data analysis, as well as serving as a guideline for the compilation and structure of the findings section of this research report. The use of the meta-evaluation framework was complemented by primary data collection via in-depth interviews with eleven respondents, drawn from SH as well as the consultancy's client base.

5.2 Sampling method

Primary data collection was based upon a non-probability purposeful / purposive sampling design, which emphasises an in-depth understanding of "…the dynamics of a particular experience or condition…" (Unrau et al 2007: 282) and thus leads to the selection of information-rich cases for study purposes. As noted by Patton (2002: 46) "Information-rich cases are those from which one can learn a great deal about issues of central importance to the purpose of the research…". The sample for this research project was therefore based upon the selection of key informants from each of the client organisations whose evaluations were selected for this meta-evaluation. Furthermore, the selection of these informants focused upon their integral role in the planning and execution of the evaluation; coupled with their possible designation as programme contact person for all communication / liaison with SH. It is proposed that the inclusion of such respondents allowed for insightful and 'rich' or detailed input as required in a study of this nature. In addition to the interviews with client organisation representatives, the four partners of SH were also asked to participate in this study. Thus the final sample for this meta-evaluation was as follows:

Table 6: Sample for meta-evaluation of Southern Hemisphere impact evaluations / outcome assessments

Stakeholder Group	Interview type	Explanation	
Southern Hemisphere	Personal, face-to-face, semi-structured interviews	SH consists of a core team of four partners, who are responsible for the strategic direction and evaluation approach of the consultancy, as well as the selection of - and tendering for - projects. Each of these partners was interviewed regarding the consultancy's evaluation model as well as the specific project/s that they managed and / or participated in.	
Client organisation representatives / key informants	Personal, telephonic / Skype semi-structured interviews	HAICU	Projects Officer: Information, Education and Communication Portfolio
		Save the Children UK	Monitoring and Evaluation (M&E) Manager: Musina Emergency Response Programme
		LEGO Foundation	Senior Vice President for Research, Development & Education
		Hands on Technologies (Implementing agent of LEGO Care for Education Project)	Director
		Isibuko Sempilo (Administrating agent for the Ponahalo de Beers Trust)	Director
		UNICEF Zambia	Child Protection Officer
		IOM Zambia (Migration Management)	Programme Officer
Total number of interviews conducted:		**11**	

5.3 Primary and secondary data collection

The following section outlines the data collection process that was followed.

5.3.1 Primary data collection

Five personal, face-to-face and six telephonic / Skype, *in-depth, semi-structured interviews* were conducted. These interview sessions lasted for approximately 60 – 90 minutes and were based upon two standardised, open-ended interview schedules, one compiled for each of the two 'sets' of respondents; namely, the SH partners and their clients.

The instrument design was undertaken to ensure a level of consistency in the questioning so as to facilitate thematic data analysis, as well as the efficient use of interviewee time. The overall aim of the interview was to obtain a high level of insight into the respondent's experiences and opinions regarding the evaluation process, its outcomes, findings regarding programme impact / outcomes, and the organisation's possible use thereof. The following section indicates some of the questions asked per criterion[32].

Credibility

Interview schedule with SH partners:

- You define your model as a 'participatory' one. What, in your opinion, makes it participatory?
- The SH model tends to focus upon the use of qualitative research methods. What quality assurance methods do you incorporate into your qualitative evaluation processes to ensure that your research outputs are credible, trustworthy and plausible?
- What sampling method/s does SH utilise? Could you explain to me why you make use of such a method/s? What, in your opinion, is the value thereof?
- Do you make use of any review mechanisms? If so, could you please indicate what these are?

Interview schedule with SH clients:

- Did the evaluation address your research aims and objectives?
- Did the evaluation address both the pre-established and any emergent information needs of your organisation? If yes, could you please provide an example of this? If not, could you please indicate why this was the case?
- Did you perceive the evaluation findings to be believable and trustworthy? Do you feel that it offered an accurate measurement / reflection of your programme's outcomes / impact?
    - If so, could you please provide reasons for your answer?
    - If not, could you please indicate why this is the case?

[32] The full interview schedules are attached as Annexures 2 and 3.

- Did you find the evaluation outputs useful?
    - If yes, could you please indicate why?
    - If no, could you please indicate why this is the case?
- Have the evaluation findings and recommendations led to programme changes of any kind?
    - If so, could you please provide me with examples of such changes?
    - What effect (if any) have such changes had on your programme / intervention?

Transferability

Interview schedule with SH clients:

- Do you believe that the results of the evaluation could be transferred / generalised to other contexts / settings?
- If so, has this been done; could you please provide examples of this? If not, could you please indicate why?

Dependability / Auditability

Interview schedule with SH partners:

- Could you describe SH's data collection process/es for this particular evaluation? Did these processes deviate from those usually applied in the SH model? If so, how?
- Please outline the data coding and analysis procedures utilised for this particular evaluation? Does this procedure generally apply to all SH evaluations? Has it been adapted at all to suit the specific requirements / demands of this particular outcome / impact evaluation?
- What were the strengths / successes of this project?
- What challenges did you experience in terms of this evaluation? How did you overcome such challenges?

Interview schedule with SH clients:

- Do you perceive the findings of the study to be consistent and coherent?
- Do you think that the analysis of the data clearly and logically emerged in the evaluation findings, conclusions and recommendations?
- Do you think that similar results would have emerged if an alternative evaluation method or a different evaluation team had been used for this study?

Confirmability

Interview schedule with SH partners:

- Please give me an overview of the instrument design process for this evaluation. Was this adapted in any way for this specific evaluation? Were the instruments reviewed? If so, by whom?
- Were the instruments for this evaluation piloted prior to commencing fieldwork? If so, could you explain how this took place? If not, could you please tell me why?
- Could you describe SH's data collection process/es for this particular evaluation? Did these processes deviate from those usually applied in the SH model? If so, how?
- Please outline the data coding and analysis procedures utilised for this particular evaluation? Does this procedure generally apply to all SH evaluations? Has it been adapted at all to suit the specific requirements / demands of this particular outcome / impact evaluation?
- Your evaluation model incorporates the use of the DAC criteria. Were these applied in this evaluation? If so, were clear definitions of the evaluation criteria formulated and communicated to your clients?

Interview schedule with SH clients:

- Did you perceive the evaluation team to be value-neutral / impartial / objective?
- Did the consultancy offer a balanced report in terms of the intervention and its successes / challenges? And in terms of representing the views of all the relevant stakeholders?

5.3.2 Secondary data collection

A project document review was undertaken focusing upon the secondary sources of evidence outlined in the table below.

Table 7: Secondary data collection as per criterion

Key Criteria	Quality Indicators	Secondary Data Sources
Credibility	1.1 Evidence of member checks / reflections on research design, process, findings and recommendations.	1.1 Evaluation inception report (where applicable), draft and final research reports generated for the evaluation, plus client interviews. Attendance registers (where possible) to indicate number and variety of stakeholders involved in planning and feedback sessions.
	1.2 Evidence of investigator and/or data source and/or methodological triangulation.	1.2 Methodology sections in evaluation proposal, inception report (where available), plus draft and final reports.
	1.3 Evidence of external review of the research process and findings.	1.3 Final evaluation report and any appendices.
	1.4 Rationale and description of sample composition and selection (for example, characteristics of sample, basis for inclusions and exclusions, sample size and how sample allowed comparative data and / or negative case analyses to be undertaken).	1.4 Proposed sample in evaluation proposal, finalised sample in inception report (where applicable), plus methodology and findings section of draft and final reports.
Transferability	2.1 Evidence of 'thick', detailed descriptions of research context to enable readers' assessment of transferability.	2.1 Background, findings and recommendations sections of final evaluation report.
Dependability / Auditability	*3.1 Dependability audit:*	
	3.1.1 Inclusion of detailed audit trail of data collection process and procedures thereof.	3.1.1 Methodology section of final evaluation report.
	3.1.2 Inclusion of detailed audit trail of data analysis process and procedures.	3.1.2 Methodology section of final evaluation report.
	3.2 Documentation of any changes to evaluation design, including reasons and any implications thereof.	3.2 Final evaluation report.
	3.3 Documentation of any study limitations and their possible impact upon the evaluation.	3.3 Final evaluation report.
Confirmability	*4.1 Confirmability audit:*	
	4.1.1 Evidence of use of multiple fieldworkers /interviewers.	4.1.1 Final evaluation report.
	4.1.2 Indication that fieldworker training was conducted.	4.1.2 Final evaluation report.
	4.1.3 Evidence of the use of appropriate and	4.1.3 Interview / focus group schedules; survey instruments (where applicable).

	unbiased questioning techniques in data collection instruments. 4.1.4 Evidence that data collection instruments were piloted. 4.1.5 Indication that quality checks were conducted on submitted transcripts. 4.1.6 Indication of the use of multiple coders for inter-coder cross-checks. 4.1.7 Use of peer review of coding and data analysis. 4.2 Use of a set of pre-defined and clearly stipulated criteria against which the programme /intervention will be assessed.	4.1.4 Final evaluation report. 4.1.5 Final evaluation report. 4.1.6 Final evaluation report. 4.1.7 Final evaluation report. 4.2 Evaluation proposal and inception report (where applicable), plus draft and final reports (the findings section will be included in the meta-evaluation to determine if the same criteria were reported against as per the definitions provided).

5.4 Triangulation of data

The use of a document review coupled with interviews with the SH partners and a variety of representatives from different client organisations, was undertaken as a means of triangulating the data obtained. Gillham (2000: 13) describes the notion of triangulation as the approaching of a research problem from different methodological standpoints; whilst Miles and Huberman (1994: 266) define it as "…to support a finding by showing that independent measures of it agree with it or, at least, do not contradict it."

Denzin (as cited in Miles and Huberman 1994) refers to a number of means of triangulating data, including the use of different data sources (such as persons, times and places), different data collection methods (including observation, interviews, or focus groups), different researchers, or by variation in data type (that is, quantitative versus qualitative). As noted above, the first means of triangulation applies to this research project.

5.5 Pilot study

Babbie and Mouton (2007) note that a pilot study is necessary to refine data collection methods, to deepen insight, and to improve upon methodological and conceptual aspects of a study. For this purpose, the instrument developed for the interviews with the four SH partners was piloted with the founding member of the consultancy to ensure that all relevant questions for the SH partners were included and that the operationalisation of key concepts had been accurately undertaken. In addition to this, the instrument for interviews with client organisation representatives was piloted with the M&E Manager of the Musina Emergency Response Programme for the same reasons as those noted above. Where necessary, adjustments to the instruments were made. In addition to this, a transcript of the

interview conducted with the M&E Manager noted above was emailed to the respondent for a 'member check' of the feedback recorded.

5.6 Data analysis

As noted by Miles and Huberman (1994: 281), "Although qualitative studies are rich in descriptions of settings, people, events, and processes, they often say little about how the researcher got the information, and almost nothing about how conclusions were drawn." In an attempt to overcome this often cited lack of rigour in qualitative data analysis (De Wet and Erasmus 2005), the following section offers a comprehensive outline of the data analysis methods and procedures employed in the course of this study.

Miles and Huberman (1994: 12) assert that qualitative data analysis is a "…continuous, iterative enterprise…", consisting of three concurrent 'flows of activity'. These include *data reduction,* where data is selected, simplified, coded, abstracted, focused, and transformed as part of on-going analysis, until the final report is completed; and *data display*, where data is compacted and presented in an organised way, so as to permit *conclusion drawing and verification.* Here regularities, patterns, explanations and causal flows emerge from the data and are tested by various means to ascertain their level of 'sturdiness', or plausibility and confirmability. The data analysis for this project will now be outlined, utilising these three 'flows of activity' as a means of structuring the discussion.

Note that, although the following description appears in a linear or step-by-step outline, the data analysis for this research project was accompanied by the repetition of steps and subsequent revision of the NVivo 10 nodes and sub-nodes[33] as well as the double-coding[34] of certain interview extracts. The iterative process of qualitative data analysis allows for such revisions as indicated by Miles and Huberman (1994: 61), who state that "Codes will change and develop as the research continues…".

5.6.1 Close reading

The first step of the data analysis process was to conduct a *close reading* of all of the interview transcripts as a means of developing an initial sense of key issues (De Wet and Erasmus 2005). Upon first reading, it was felt that the client organisation representative feedback had been, overall, markedly positive. However, a second close reading started to reveal some concerns amongst the respondents, specifically regarding methodological issues.

[33] NVivo 10 software uses the term 'nodes' as opposed to 'codes' and 'sub-nodes' as opposed to 'sub-codes'.
[34] 'Double-coding' is used here to refer to the entry of a particular section of interview transcript text at more than one node or sub-node.

5.6.2 Data reduction

Step 1: Following the close reading of interview transcripts, a tentative *list of descriptive and interpretive codes*[35] was developed for data analysis purposes. This list was based upon a consideration of the central research question as well as the meta-evaluation criteria listed in the analytical framework. These listed codes and sub-codes were then entered as 'nodes' and 'sub-nodes' in the NVivo 10 software programme, which I elected to use as a means of assisting with data categorisation and organisation[36].

Step 2: The next step of data reduction was to code the data. All interview transcripts were saved electronically and then imported into the NVivo 10 software programme for the computer-assisted coding process. Coffey and Atkinson (1996, cited in De Wet and Erasmus 2005: 30) note that "…coding helps one to organise, manage, interpret, and retrieve meaningful segments of data," whilst simultaneously allowing for a level of analysis in that segments of text are being 'matched' to a particular concept/s. Miles and Huberman (1994) refer to this process as *first level coding*.

Each interview transcript was read and coded in its entirety before coding the next. As noted by De Wet and Erasmus (2005), this system helps to maintain the 'integrity' of each transcript in that it prevents the respondents' voices from merging with one another in the mind of the researcher. The coding process began with the four SH partner interview transcripts before moving onto the client respondent's input.

It should be noted that this process was greatly facilitated by the formulation of an operational definition for each of the selected meta-evaluation criteria, many of which were simply transposed as a code or sub-code. This enabled me to check whether or not the text segments that were being coded corresponded with the definition of that specific criteria / code, thus facilitating a more rigorous analysis.

Step 3: Following the above two steps, *second level coding* was conducted, which includes 1) the identification of clusters and hierarchies of information, and 2) a "…deeper level of analysis…" including the identification and classification of emerging data themes or patterns, referred to as *pattern coding*[37] (De Wet and Erasmus 2005: 33). Pattern coding has four important functions (Miles and Huberman 1994: 69); namely:

[35] Miles and Huberman (1994: 56) define codes as a data labelling device; that is as "…tags or labels for assigning units of meaning to the descriptive or inferential information compiled during a study…", where descriptive codes attribute a class of phenomena to a segment of text; while interpretive codes are used to draw distinctions within the descriptive codes and categorise them accordingly.

[36] The NVivo 10 software programme allows users to organise 'chunks' of information according to its nodal system of categorisation. It is not used to perform any form of analysis of entered data, but rather allows the research to find, cluster and extract information segments relevant to a particular research theme / topic or question.

[37] Pattern coding is defined as a "…way of grouping summaries together into a smaller number of sets, themes or constructs…" as a means of looking for common themes or threads that group different data sets together (Miles and Huberman 1994: 69).

a) It reduces large amounts of data into a smaller number of analytical units;
b) It starts the analysis process by allowing for the emergence of certain themes or trends that could be explored further in any fieldwork that may still be on-going;
c) It helps the researcher to construct a 'cognitive map' to facilitate understanding of emerging themes or trends; and
d) For multi-case studies such as this one, it facilitates cross-case analyses and, hence, the drawing of conclusions.

For example, the second level coding conducted for this study 'surfaced' a variety of methodology issues, which were discerned both across codes and across the various evaluations conducted by the consultancy. A second level code *Methodology* was thus created as a means of clustering data sets pertaining to SH's methodological choices and the clients' perceptions thereof. This was extremely useful as methodological issues were highly pertinent to the central research question and to findings on client perceptions regarding the accurate measurement of development intervention outcomes and impact.

5.6.3 Data display

Following the coding process on NVivo 10, electronic reports on each of the first level codes / nodes and sub-codes / sub-nodes were generated and printed. These summaries provided a number of appropriate and descriptive quotes for the final report, plus allowed for the weighting of evidence based upon the number of times a specific issue was mentioned by the various respondents. These coding reports were used in conjunction with an Excel spreadsheet meta-evaluation summary, another means of data display, which was compiled following the document review (referred to in section 5.3.2 above)[38].

5.6.4 Conclusion drawing

The use of the NVivo coding reports and Excel spreadsheet / meta-evaluation summary allowed for a structured and comprehensive analysis of both primary and secondary data sources. Key themes, relating to each of the four quality criteria, that emerged from the interview data were cross-checked against the same themes in the meta-evaluation summary. Conclusions were drawn for each criterion based on areas of convergence and divergence between the data sources.

5.6.5 Data verification

As previously noted, the triangulation of different data sources was one of the methods employed in this study for verification purposes. Input was obtained from both consultancy members as well as their clients, whilst the information that was obtained via the interview transcript analysis was cross-checked against data that emerged from the document review of a number of key organisational sources. In addition to this, a feedback meeting was held with the four SH partners to share the preliminary study findings and accompanying recommendations.

[38] This is attached as Annexure 1.

5.7 Ethical considerations

The following ethical considerations are noted with regard to this study:

5.7.1 My personal position

As previously noted - I am a consultant for SH and have been employed by the organisation for a period of three years. A considerable measure of reflexivity - and on-going dialogue with a skilled supervisor - was therefore required to ensure that my position did not, in any way, hamper my objectivity and judgement when assessing the credibility, transferability, dependability and confirmability of the consultancy's model employed for the purpose of impact evaluations / outcomes assessments. In addition to this, it should be noted that SH prides itself upon being a learning organisation and therefore encourages critical self-reflection based upon client feedback, to allow for learning and future improvement. As such, the consultancy has endorsed this study and its members have noted that they will in no way discourage me from following any particular outcome, even if this reflects negatively upon their work in some manner.

5.7.2 Dissemination of study findings

Secondly, one of the clients of SH indicated that she wished to receive a copy of this report. Given the findings, this request could possibly jeopardise the consultancy's future employment with this client. As noted by Hedler and Gibram (2009: 221), "A negative (meta) evaluation can harm the credibility and merit of a given institution or group of evaluators." Therefore, it was elected to raise this issue with the SH partners during the feedback and recommendations session and to formulate a response to this request in conjunction with the consultancy's members.

5.7.3 Participant consent and anonymity

Finally, it should be noted that all client organisation representatives and SH partners were contacted via email to request their written consent to participate in this research project. In addition to this, their verbal consent to participate in this study was obtained prior to commencement of the interview (Please refer to the introductory section of the client and SH interview schedules). Client anonymity[39] was assured during this contact process; in addition to which it was confirmed – via email correspondence – that only the evaluation project would be indicated in this report and that only the client's professional title would be noted on the sample list.

5.8 Critical reflections on methodology

The following observations were noted in the course of this research study:

a) The completion of interviews with the SH partners prior to those with the client organisation representatives greatly facilitated critical reflection regarding the meta-

[39] Anonymity refers to the "…lack of identifiers…" or any other form of information that would indicate which individuals or organisations provided certain sections or segments of data (Miles and Huberman 1994: 293).

evaluation criteria and the formulation of appropriate questions for the client interview schedule.

However, a number of areas for future improvement were noted, including:

b) I should have confirmed client organisation representatives' availability to participate in the study prior to formulating my instruments and interviewing the SH partners. One of the evaluation projects selected for this meta-evaluation had to be eliminated from the study as the relevant staff member had - in the interim - left the organisation, leaving no-one with the necessary level of institutional memory to participate in my research project. Consequently, another project had to be selected and reviewed, including additional interviews with the SH partners involved with this project.

c) SH's definition of impact was clarified via the document review and during conversations with each of the four partners. It was also noted during discussions with the consultants that SH's definition of impact was shared with client organisation representatives during the planning meeting and a common understanding of the term - for evaluation purposes - was established in this manner. However, the client respondents were simply asked if they believed that the evaluation project undertaken by SH had offered an accurate and valid measurement of impact. Their understanding of the term should have been probed into in a far more precise manner; particularly in light of the central importance of this concept to this research project.

d) The data analysis could have been improved upon in terms of a higher level of reflection on my part, specifically in terms of the use of certain codes and my assignment of segments of interview text to such codes. For example, many of the codes in the final coding reports included data that was purely categorical as opposed to interpretive. In other words, text segments were coded according to the question that was asked as opposed to critically reflecting upon the answer that was offered. This meant that a substantial amount of double-coding was needed, which considerably lengthened the coding reports and hampered data analysis.

Chapter 6 Findings

The following section offers a discussion of the key findings relating to this study. These will be discussed with reference to the central research question and as per the meta-evaluation criteria noted in Table 5 of Section 4.2. The discussion is based upon an analysis of the Excel spreadsheet meta-evaluation summary, attached to this research project as Annexure 1, as well as an analysis of the NVivo 10 primary data coding sheets.

6.1 Credibility

When assessing the credibility of SH's evaluation model, evidence of member checks, triangulation methods and external review mechanisms, was sought. In addition to these quality indicators, the specification, rationalisation and use of a clear sampling framework, was also used as a means of verifying credibility; along with levels of perceived trustworthiness, 'believability', relevance and usefulness of evaluation findings and conclusions amongst client respondents.

6.1.1 Evidence of member checks:

A review of secondary data sources, including the evaluation proposal; inception, draft and final reports; as well as any relevant meeting agendas and PowerPoint presentations, indicates that the SH evaluation model is of a highly participatory nature and thus provides a number of opportunities for member checks or member reflections on the data and the analysis thereof. Four key member checks or member reflection opportunities were discerned; namely:

- An evaluation planning workshop or planning meeting, attended by client organisation representatives, where the scope and expected outcomes of the evaluation are clarified and agreement obtained regarding key issues to be addressed, sample frameworks, timeframes and evaluation methodology;
- The submission of evaluation instruments or interview schedules by SH to the client organisation for review and comment;
- The submission of a draft report to the client organisation for distribution, review and comment; and
- The hosting of a feedback and recommendations workshop, including members of the client organisation as well as a variety of other, relevant stakeholders, where evaluation findings are presented and discussed, followed by the 'workshopping' of recommendations.

The latter two forms of member check are specifically noted by Lincoln and Guba (1985) as a strategy to enhance the trustworthiness and credibility of one's research; while Patton (2008) supports these processes in his assertion that one of the key means of enhancing one's research validity is to provide key informants with the opportunity to review and

correct the outputs thereof. This would have taken place via the distribution of the draft report for comment, as well as via the hosting of a feedback workshop.

Primary data collection also confirms the participatory nature of the SH evaluation model as well as the provision of opportunities for client organisation 'checks' and input. All client organisation representatives indicated that they perceived the SH model to be not only participatory, but also highly collaborative. This was noted by all seven of the client respondents as a key strength of the SH model; whilst one respondent indicated that – in her experience of evaluation practice – it was also fairly unique:

"In fact, I think that the process was quite unique in that we collaborated with SH on so many different aspects and this collaboration was on-going. I think that the level of engagement was very high and yes, the whole process was very much participatory…it was very much a joint effort, with a lot of discussion and consensus between ourselves and SH." (Client respondent A)

However, it should be noted that the member checks referred to above were generally conducted solely by members of the client organisation or affiliates thereof – and not by development intervention target group members or beneficiaries. I would argue that in order to be deemed a truly participatory model – and to confer a higher level of integrity and credibility on the research outputs of the SH evaluation process – member checks amongst all key respondents in the research process should be prioritised. This observation was noted during my conversations with the SH partners and, while all agreed with this assertion, it was noted that invitations to attend the feedback and recommendations workshop were at the client organisation's discretion. While SH could make suggestions as to who might be included at such a member check, resource and time constraints often prevented the attendance of those outside the client organisation. Nevertheless, the budgetary inclusion of a feedback session with target group and beneficiary stakeholders in future evaluation study proposals is an important consideration for the consultancy to take forward.

6.1.2 Evidence of investigator/researcher, data source, data collection method and/or methodological triangulation:

The secondary data source review indicates that all of the SH evaluations included in this meta-evaluation included the triangulation of data sources and data collection methods, as well as investigator / researcher triangulation.

For example, the *Proposal for the Evaluation of Save the Children UK's Response to the Situation in Musina since 2008* (June 2010: 7) notes that a qualitative methodology would be utilised for the study; combining a literature and project document review, a stakeholder workshop with participants who work across relevant sectors, "...to allow for them to validate each other's responses and to balance perspectives..." (2010: 9), focus group sessions with beneficiaries (children), and in-depth interviews with Save the Children UK's

key project personnel and relevant programme partners plus with government officials involved in the child protection sector. This indicates that both data sources and data collection methods could be triangulated in this study, which was confirmed by a review of the reporting in the findings sections of the draft and final reports for this evaluation. For example, the findings section of the *Final Report for the Evaluation of Save the Children UK's Response to the Situation in Musina since 2008* (September 2010) includes comparative analyses as per stakeholder group as well as appropriate quotes to support the arguments made. The following quote refers: "For example, SCUK's international and local partners hold the general view that there is effective coordination of services in Musina. However, SCUK staff members state that this has not always been the case..." (2010: 33); and "All of the interviewees agreed that the Musina Programme has been successful in raising the government's awareness of issues related to children on the move...". (2010: 42). Similar observations were made following a review of the proposals, draft and final reports of three other evaluation studies included in this meta-evaluation. Only one of the studies' final reports did not include any specific indication of data source or data collection method triangulation; that is, for the *Evaluation of the Ponahalo De Beers Trust Programme*.

Methodological triangulation was included in only three of the five evaluations due to the sole use of qualitative research methods in two of the studies under review[40]. However, where both qualitative and quantitative methods were used, triangulation of the data obtained was evident. An example of this, as obtained in the *Final Report: Research Study of Student Risk Behaviour and Evaluation of ACES Programme for HAICU* (July 2009: 117), is as follows: "Another prominent reason for not getting tested is fear of knowing one's status, cited by 21% of residents. This finding was supported in the qualitative component of the research – and, ironically, was expressed by both people who were low risk, i.e. not sexually active and high risk, i.e. those who were sexually active."

The consultancy's efforts to incorporate triangulation methods as a means of enhancing the credibility of its work were also noted during the interviews conducted for this research project. All four SH partners noted that, despite having a reputation as a predominantly 'qualitative research' consultancy, they had begun to incorporate the use of mixed methods into their evaluation practice more frequently. This was due to a number of factors, including a rise in client demand for a mixed methods approach plus the perceptions of scientific rigour that tend to be associated with the use of quantitative research. However, all of the partners noted that the use of mixed methods was particularly valuable as it allowed for triangulation of their data, which added depth and quality to their work. The following quote refers:

"Our model also promotes the use of mixed methods; that is both quantitative and qualitative methods; as well as looking at a single issue from different points of view…triangulation…". (SH partner 3)

[40] The remaining three evaluations incorporated the use of a mixed methodology.

6.1.3 Evidence of external review of research process and findings:

Three of the five evaluations reviewed included some indication that the study and/or its outputs had been subject to external review. These are individually noted below:

a) An internal work plan for the *Research Study of Student Risk Behaviour and Evaluation of ACES Programme for HAICU* (2009) notes the use of an external expert on quantitative research to guide and inform this aspect of the study.
b) The proposal, progress, draft and final reports for the *Outcome Assessment of the Care for Education Project* for LEGO note that the interview schedules were developed with assistance and input from subject matter experts based at the Psychology Department at the University of Cape Town (UCT) and the Early Learning Resource Unit. These documents also note that comments on the evaluation process and instruments were received from a pedagogic psychologist based at LEGO Serious Play, while Annexure 1 to the Final Report, titled *Qualitative endorsement: Final Report on the Outcome Assessment of the Care for Education Project: Southern Hemisphere Consultants*, compiled by an external subject matter expert, offers an overview and endorsement of the study and its findings.
c) The *Final Report for the Evaluation of the UNJPHT* (21 December 2012: 10) notes that a draft report was submitted to an international, human trafficking expert for comment. This feedback - it is noted - was included in the final report. A footnote cites the expert advisor's qualifications as a Masters' Degree in International and Intercultural Management, a Postgraduate Diploma in Non-governmental Organisation (NGO) Leadership and Management, plus 10 years of experience in the field of migration and human trafficking.

These findings were corroborated during the interviews with SH partners; that is, it was noted during these discussions that the consultancy often employed the services of an external, subject matter specialist to inform a study and review / comment upon its results. This measure was not undertaken solely for research credibility purposes, but also because the SH partners feel that they are not subject specialists in all the sectors in which they are employed to conduct evaluations. Hence, to ensure a high quality service to their clients, expert opinion is sought by the consultancy as and when necessary. This is outlined in the quote that follows:

"Another part of the model is that we acknowledge - in each and every sector that we move into - that we don't have all the necessary expertise. In the case of the X evaluation, for example, we were dealing with early childhood development in a developing country context. So we had to have a lot of input from experts…". (SH partner 2)

Feedback from the client respondents appears to indicate that the use of external subject experts enhances the level or perceived trustworthiness of the study results. Two of the respondents included in this study specifically noted that the employment of a subject

matter expert had contributed towards their perceptions of the evaluation's credibility. The quotes below illustrate this further:

"What we also appreciated was that SH engaged the services of an expert on human trafficking to give that extra input and insight. The core skills were there in the team, but then to get the expert involved; that was a good idea and we definitely appreciated this." (Client respondent G)

"So yes, in terms of offering an accurate measurement of outcomes and impact, absolutely; I do not doubt it. Remember that we had a statistician look at it, …assisting us with the sample…Given this, I believe that the research was credible…". (Client respondent A)

6.1.4 Rationale and description of sample composition and selection:

The findings for Quality Indicator 4 of *Credibility* offer a somewhat mixed result. Three of the five evaluations reviewed included an explicit sampling method and, similarly, three of the five evaluations outlined a definite rationale for the selected study sample. While all five of the evaluations included a clear outline of the final sample for each study, one of the evaluations did not include any analysis or reporting across the specified study's sample groupings.

It was also noted that little input was offered in general in the evaluation project documents regarding the selected sampling method and implications thereof in terms of potential bias in evaluation findings. This was particularly evident in discussions regarding qualitative sampling frames. For example, the *Proposal for the (Impact and) Outcome Assessment of the Care for Education Project: "Developing Talents through Creative Play" in Atteridgeville Township* (2010:10-11) includes a table outlining the sample selection process for each of the study's respondent groups as well as an explanation/rationale for the employment of the sampling technique. Similar detail is noted in the draft and final reports. However, where it is stated that a random sample of five schools was taken from 11 schools in and/or serving the Atteridgeville area, in conjunction with the implementing agents' input, there is no indication as to how the agents' involvement in the sample selection process may have generated a level of bias.

It was noted in the course of the interviews with the SH partners that methodological input in the final report is often limited as per client request. It was reported that clients often imposed reporting constraints upon the consultancy, asking for brevity and inclusion of only the most salient points. Given the emphasis on results-based evidence, as previously discussed, the emphasis of any evaluation report tends to be upon the findings, conclusions and recommendations. This relegates methodological input to a small portion of the final document. However, I would argue that such a lack of methodological explicitness detracts from the credibility of a study and certainly requires reflection on how best to address the situation in the future. A means of overcoming the restrictions to final report length is to possibly include a clear methodological account as an annexure to the main document,

which the client organisation may choose to share upon dissemination of the evaluation results.

6.1.5 Client feedback indicating that SH evaluation outputs are perceived as trustworthy, credible and believable:

A high level of trust and confidence in the evaluation findings was voiced by all of the client organisation representatives. It is argued here that the high level of engagement and participation facilitated throughout SH's evaluation process contributes towards such high levels of perceived credibility and trustworthiness. This is confirmed in the quote that follows:

"I think that the nature of the process – how the outcomes came about – this certainly contributed to the credibility of the study." (Client respondent G)

However, it might also be argued that the participatory and utilisation-focused approach, with its high level of client collaboration and input, does imply a level of bias in that the evaluation is substantially informed by the clients' own, specific value system and developmental perspective; thus presenting little risk of any form of alternative epistemological framework. As noted by Gomm (2008: 329) "…evaluation researchers face a problem of value pluralism, which raises questions about whose values, desires, preferences and so on should inform the criteria for evaluation." In the SH evaluation model, the client organisation informs the scope, criteria, instrument design, findings and recommendations of the evaluation. Furthermore, as noted by Bamberger and White (2007: 66), evaluation is often motivated by development agencies' need to secure on-going funding. As a result, there is a considerable degree of pressure on the evaluation consultancy to ensure that the results are positive. The inclusion of member checks in the evaluation process then has a *detrimental effect* upon an evaluation's trustworthiness and credibility, plus makes the reconciliation of evaluation results with the needs of the users a complex and politicised task.

Therefore, I would argue that it is the participatory nature of the SH evaluation model that both strengthens *and* weakens the credibility thereof. Of note is that three of the client respondents, whilst indicating that they appreciated SH's participatory approach, simultaneously indicated that they would have preferred SH to take more 'ownership' of the evaluation process - and of the results. This is elaborated upon by the two quotes that follow:

"But, overall, I think that SH had the right idea…perhaps our environment was a bit intimidating? SH had the template to follow and they did that well, but there was not enough initiative or 'push' from SH. I feel that their ownership of the final product…of that report…should be more. You know, they should tell us what they advise or want and then we can do it." (Client respondent D)

"In the definition phase or first phase we had a long discussion on how to measure impact and so we agreed to relate the impact or to measure it via the use of some of the brand values. But this led to lower credibility that we would have hoped to have gained…So X were very keen on doing the evaluation in this way; I am not blaming SH on this, but perhaps a research or evaluation expert should have told us this…you know, 'yes, this is interesting, but it is a risky way to go and you can affect your credibility'." (Client respondent C)

Conversely, Stufflebeam notes that if people inside the evaluation are confident with the evaluator and his/her findings, "…people outside the system may think that the evaluator has been co-opted…" (Stufflebeam 1974 / 2011: 115). This was, indeed, noted by three of the respondents who indicated that their peers had felt that the SH evaluation had rendered an overly positive result. As one respondent observed:

"Input from other staff was that it was an overly positive evaluation and that it did not adequately represent the situation." (Client respondent B)

These findings would suggest that the use of a participatory approach - and the implications thereof in terms of perceived credibility of evaluation outputs - are issues that require a measure of reflection (and address) by the consultancy in the future. The following suggestion was offered by one of the client respondents:

"I think that it (SH model) compares well (to other evaluators), but one bit of advice that I would like to offer is that SH should be more 'forceful' or firm in their approach, methods and with regard to the evaluation findings. Perhaps when negotiating a contract with a client/s, the consultancy should think about setting parameters as to where feedback offered can - or cannot - be taken. There should be a level of reporting where the client should only have limited input and which they cannot influence with their own political agenda." (Client respondent B)

6.1.6 Client feedback indicating that SH evaluation outputs are perceived as relevant and useful, incorporating supportive evidence:

Given that this study's definition of credibility included the perception amongst stakeholders that the evaluation evidence produced was also *useful* and *relevant;* the client respondents included in this research project were asked whether they had, indeed, found the evaluation findings useful and, if so, whether they had utilised the findings and recommendations for some form of programmatic and/or organisational change.

All of the client organisation representatives that participated in this study indicated that the SH evaluation had addressed - and been relevant to - their information needs upon commencement of the evaluation as well as those information needs that had emerged over the course of the study. As noted by one client:

"Yes, it certainly did meet our needs. It helped in that we had a very clear Terms of Reference that we had thought through beforehand. This was used to guide the process…but in

addition to that, the process was very participatory and so adjustments could be made and jointly decided upon as and when necessary." (Client respondent G)

As noted by Patton (2002: 184) "…people who participate in creating something tend to feel more ownership of what they have created and make more use of it." This assertion was certainly supported by the feedback obtained from SH's clients, who all noted a variety of examples indicating both 'process use' as well as 'process impacts' (Patton: 189)[41]. Five respondents noted that programme / intervention changes had been made based upon the findings and recommendations included in the consultancy's evaluation report[42]. In terms of the possible impact of such changes upon the programme's effectiveness, three of the respondents noted that the changes had only recently been implemented and that it was, therefore, too early to speak of any possible effect. However, one of the respondents noted the following:

"…we did use the recommendations and they have certainly added value to the programme…the programme is going from strength to strength – we can see the benefits of these changes. The impact has been huge. In fact, it is now being referred to as the 'lighthouse project and Y wants to expand it to other, similar contexts. So yes, you can certainly tick all of the recommendations." (Client respondent D)

When asked if the evaluation findings and recommendations had been used for organisational development purposes, six of the seven client respondents noted that the evaluation outputs had been used for this purpose. Examples offered of relevant organisational changes included improved budgeting processes to enable better M&E systems and enhance beneficiary representation on the programme; the development of an internal M&E resource for the planning and implementation of future evaluations; the investigation of higher levels of engagement and cooperation with affiliated multilateral organisations; the improvement of organisational planning and operation; and the development of improved human resource management measures. Again, the respondents felt that some of these measures had been too recent to report upon the effect thereof. However, those that reported such changes were all of the opinion that they would lead to improved effectiveness and performance on the part of their organisation.

Davidson (2005: 213) notes that all evaluation design decisions involve trade-offs; for example, increasing stakeholder involvement does improve upon utility, but decreases perceptions of credibility to outside audiences. However, he adds "…these drawbacks are

[41] 'Process use' is defined by Patton (2002: 189) as those changes in thinking and behaviour that take place amongst those individuals that had been involved in the evaluation process, as a result of the evaluation process; whereas the same author defines 'process impacts' as including those changes in a programme or in organisational procedures, culture and systems, that may emerge due to the evaluation process.

[42] The remaining two respondents noted that funding for Phase 2 of the intervention had not yet been received hence no programme restructuring had taken place at the time of their interviews. However, both respondents indicated that such a future programme would certainly be based upon the findings and recommendations offered by SH.

minor in comparison with the substantial gains in organisational learning capacity that are gleaned from a well-conducted participatory evaluation." This might be argued as being the case for SH's evaluations. The level of perceived relevance and usefulness was high, and all seven of the client respondents noted a high level of gain in terms of organisational learning and capacity development.

6.1.7 Conclusion

Overall, the SH evaluation model fares well in four of the six quality assessments for the criterion, *credibility*, with mixed findings for the remaining two quality indicators.

A highly participatory process, the SH model allows for member checks / member reflections as well as a high level of researcher, data source and data collection method triangulation. Furthermore, it appears as though the consultancy is moving towards an increased use of mixed methods, thus enabling an increasing level of methodological triangulation as well.

The use of external reviewers was apparent in three of the five evaluations, while feedback from the SH partners indicates that they are open to the use – and input – of external subject matter experts as and when necessary. This willingness to include the input of external reviewers was noted by the client respondents as something that they appreciated - and which they believed contributed to the credibility of the consultancy's work in the evaluation and development sectors.

Clear sample frames were included in all of the evaluation reports and four out of the five evaluations included extensive analysis and reporting as per the specified respondent groupings, thus supporting the use of purposive sampling methods. However, there was erratic input in terms of sampling methods as well as the provision of a clear rationale for selected sampling methods. This was particularly evident in qualitative studies or in the qualitative component of mixed method research.

Finally, client respondents were unanimous in declaring SH's evaluation outputs relevant, useful and credible. However, further analysis of client input revealed some concern around the consultancy's level of ownership of the evaluation process, while concerns were also raised regarding the level of perceived co-option, both amongst the clients themselves as well as their external audiences. Thus, while the participatory approach was perceived as a positive attribute of the model, it was also acknowledged as contributing towards diminished levels of perceived credibility.

6.2 Transferability

When assessing the transferability of the SH evaluation model outputs, the applicability or 'fittingness' of evaluation findings, conclusions and recommendations to other contexts or programmes of a similar nature was measured. Quality indicators included evidence of 'rich' or detailed descriptions of the research context to enable the readers' assessment of

transferability plus client feedback regarding their perceptions of the transferability of the SH evaluation results.

6.2.1 Evidence of 'thick', detailed descriptions of research context:

All of the draft and final reports, generated by SH for the evaluations under review, included approximately five to six pages of background input regarding the context of the programme under evaluation as well as detail regarding the programme - and the evaluation - itself. For example, the *Final Report for the Evaluation of the Ponahalo De Beers Trust* (20 April 2012) includes a six page discussion detailing the following: the background to the PDT, including the history, aims and objectives thereof; and a description of the geographic sphere of activity for the PDT interventions, including the socio-economic environment around the Finsch and Voorspoed Mines (where the projects under review are situated). This is followed by an overview of the projects themselves, including organisational structure, aims, activities, and target groups; and finally, an outline of the scope of the evaluation. Similar levels of detail were noted in the final reports of the other four evaluations.

Hence, it is argued that, overall, the background input provided offers a succinct but clear contextualisation of the programme/s under evaluation as well as a detailed description of the programme/s themselves. Such input would enable the reader to establish the transferability of the evaluation findings and recommendations to other, similar situations and/or similar programmes.

6.2.2 Client feedback regarding transferability of SH evaluation outputs:

When assessing client perceptions of the transferability of SH's evaluation outputs, the following questions were asked: *"Do you believe that the results of the evaluation could be transferred / generalised to other contexts / settings?"* and *"If so, has this been done; could you please provide examples of this? If not, could you please indicate why?"*

All seven of the client organisation respondents[43] interviewed indicated that, while there were certain programme-specific outputs, SH's evaluation findings could be transferred – and applied – to other, similar programme settings. Three of the seven respondents indicated that this had already been done. The following quote expands upon this feedback:

"Yes, there were findings that could be transferred to other programmes; in fact, there actually are some that are being utilised in other programmes…for example, the DRC border and Pakistan programmes. There are specific contextual issues to any programme, of course, but there are also areas of commonality…so yes, we have used several of the findings relating to these situations in other programmes." (Client respondent B)

[43] It should be noted that all of the respondents in this study were involved in the planning and implementation of a number of different projects, either nationally and / or internationally; and so would be able to answer such a question based upon their personal, professional experience.

Stame (2010: 380) argues that the greatest attention should be on the external validity of a study or the *ability to generalise results from one setting to another* (my italics). Thus, she argues, evaluation findings that can only be utilised in similar programme settings, with identical characteristics, are of little or no use, especially given the global nature of development today, with its increased levels of cooperation amongst international and national development agencies. This assertion is not without merit. However, it might also be argued that development interventions that - once evaluated - offer best practice models or that are achieving marked success within a particular context have sufficient merit within their basic logic so as to allow for their successful and useful transfer, adaptation and contextualisation in other settings. It was noted, during my interviews with the four SH partners, that the SH evaluation model often incorporates the use of case studies for such a description of best practice models[44], which would further enhance the transferability of their evaluation findings. This assertion of the model's favourable assessment against the criterion of transferability is supported by the feedback obtained from all seven client respondents included in this study.

6.2.3 Conclusion

The SH model was favourably assessed against the criterion of transferability. This finding was based upon evidence sourced via the project document review as well as input obtained from SH's clients.

6.3 Dependability / Auditability

The criterion of dependability refers to the *consistency, coherence, logic and stability of the study process and data analysis over time and across researchers and methods*. A number of quality indicators were developed as a means of measuring or assessing the SH model's level of dependability and audibility. These are individually reported upon below.

6.3.1 Dependability audit:

The dependability audit focuses on the inclusion of a detailed audit trail of a) *data collection* procedures and b) *data analysis* processes. A number of secondary data sources were reviewed to establish if any evidence of such audit trails was available. These data sources included fieldworker reports (where applicable) as well as the methodology section of draft and final evaluation reports. The following findings are noted:

Project documents of only two of the evaluations reviewed include detailed feedback regarding data collection processes. That is:

- The project documents for the *Evaluation of the United Nations Joint Programme on Human Trafficking* (UNJPHT) included a fieldworker report template and four fieldworker reports, prepared and submitted by each of the qualitative fieldwork teams. Included in the fieldworker reports is the following input: a summary of all

[44] The evaluation of the PDT included the preparation of case studies for each of the projects under review.

day-by-day activities undertaken by the fieldworkers over the course of the study; the details and a brief overview of each of the respondents that participated in the study; general observations and reports regarding the time spent in field; any limitations that were experienced and the impact thereof upon the fieldwork process and data collection; preliminary findings; and acknowledgements.

- A *Progress Report for the Outcome Assessment of the Care for Education Project: "Developing Talents through Creative Play" in Atteridgeville Township*, dated 03 November 2011, offers an overview of the fieldwork process (actual versus planned) for the period 25 - 29 October 2010. Also noted are the successes and challenges of fieldwork, key preliminary findings as per the three evaluation criteria, and a work plan for the way forward. A fieldworker report was also sourced in the project documents, dated 02 November 2010. Noted are the number of interviews and observations conducted by the fieldworker, as well as challenges and successes experienced during the fieldwork process and key findings.

However, none of the proposals or draft and final reports for the five evaluations under review included detailed input regarding data collection processes.

Findings regarding the detailing of data analysis processes and procedures were similar in that the project proposals, inception reports (where applicable) and draft and final reports included mentions of systems or software applications, such as NVivo 8 or NVivo 10, utilised for data coding purposes. However, further input – particularly regarding specific qualitative data analysis procedures – was limited across all of the evaluations included in this study. A typical example is the *Proposal for the Impact Evaluation and Strategic Planning of the ACES Peer Education Programme* (25 September 2008: 10-11), which notes that data will be analysed using multi-variate analysis techniques and according to gender, year of study, race if South African, and foreign student if not South African. It also notes that qualitative data analysis would be conducted with the assistance of NVivo 7 software while SPSS would be used for the quantitative data analysis. No further input was found in the project's draft or final report documents.

Interestingly, this finding regarding a lack of data analysis input - particularly pertaining to qualitative research methods - is echoed by one of the client respondents, as indicated below:

"Also in the methodology section of the report, I felt that it wasn't entirely balanced. There was less input given regarding the qualitative methodology and data analysis as opposed to the quantitative research methods which were very detailed. The qualitative definitely received less attention if I could describe it like this and I think it would have been nice to have more detailed input in this regard." (Client respondent G)

Conversations with the consultancy's partners indicate that despite this lack of detail in key project documents, the members of SH do incorporate the use of specific data collection

processes as well as systems and procedures for data analysis. Noted was the consultancy's use of software programmes, such as NVivo, plus analytical frameworks or coding sheets, which detail the possible codes and sub-codes for data coding and analysis purposes[45]. The following quote refers:

"…we have been mindful of developing a strong framework to use for our analysis. This we have used for the past three years, but this will also depend upon who the client is and the nature of the evaluation. I think this is important…to develop an analytical framework for the proposal or the inception report and then this indicates how we are going to be conducting the data analysis…" (SH partner 1)

This assertion was supported by one of the consultancy's clients, who stated:

"We also developed 'themed' areas and then we matched the questions to the particular theme. So, the data analysis had a definite framework, which we developed together with SH, and this facilitated the logic of the analysis a great deal." (Client respondent A)

However, no such analytical framework was explicitly referred to – or described - in the project documents of the five evaluations under review.

6.3.2 Documentation of any changes to evaluation design:

None of the evaluations included in this meta-evaluation included the documentation of changes to the original evaluation design. This is understandable, however, given that all of the studies appeared to be fairly stable and consistent throughout the research process. No major, undocumented deviations from the evaluation designs noted in the relevant evaluation proposals were discerned.

6.3.3 Documentation of any study limitations and their impact upon evaluation findings:

All five of the reviewed evaluations included a section on study limitations. Three of the evaluations offered some reflection on the impact of these limitations on the data collection process and research findings; for example, the following is noted with regard to study limitations for the *Research Study of Student Risk Behaviour and Evaluation of ACES Programme for HAICU* (2009); namely that:

- Difficulties were experienced with finding a suitable, on-campus venue for the focus group discussions with non-residence students. Consequently, an off-campus venue had to be sourced and this negatively affected attendance at the focus group discussions, despite incentives being offered. (Noted in final report is that some focus group discussions had to be cancelled).

[45] Such coding sheets were found for both the Evaluation of the Ponahalo De Beers Trust and the Research Study of Student Risk Behaviour and Evaluation of ACES Programme for HAICU. While no coding sheets were found in the documents pertaining to the other three evaluations included in this study, detailed coding reports were found in all five evaluation project document folders. The data coding and analysis framework for the Evaluation of the Ponahalo De Beers Trust is attached to this report as Annexure 4.

- Student participation was voluntary and all possible participants were informed regarding the nature of the topics to be covered. This may have attracted those with an interest in the subject matter of sexual behaviour and HIV / AIDS, thus limiting variation in input. However, the Final Report for this evaluation (pp15-16) notes that this was compensated for by the use of the quantitative sample.

The remaining two evaluations, whilst noting the limitations affecting the study, did not address the impact of such issues on the data obtained. Hence, the overall impression – formulated following this document review – is that a heightened level of reflection and input regarding study limitations is necessary to raise the level of projected dependability of SH's evaluation model.

Interestingly, once again, this lack of input regarding the potential impact of study limitations was noted by one of the client respondents. The quote below elaborates on his response:

"Also during the implementation phase…there was the teacher strike and this impacted on the timeframes allowed. The interviews now had to be done according to this limited timeframe and so this led to some doubt as to whether or not the data gathered under these circumstances would have been sufficiently up to snuff? I believe that here, as a research expert or evaluator, SH could have advised us accordingly." (Client respondent C)

6.3.4 Evidence that client organisations regard the SH evaluation outputs to be stable and replicable across researchers and methods:

When assessing client perceptions regarding the dependability and replicability of the SH evaluation model, all seven of the client organisation respondents indicated that they felt that the evaluation results were both replicable and dependable. This general sentiment is elaborated upon below, by one of the client organisation respondents:

"On the whole, we felt that we were very happy with the evaluation and the report; so in theory if someone had also conducted this evaluation in a comprehensive way the results should have been the same. Mixed methods and the triangulation of data that this allowed…this also validated the outputs of the evaluation. I believe it was both highly credible and reliable." (Client respondent G)

One of the respondents noted that it was common practice for her organisation to offer evaluation data for peer review. This had been done during the SH evaluation and the feedback obtained from the external consultant (based upon their analysis of the raw data submitted) had been identical to that of SH. The client respondent noted this favourable peer review as a means of justifying her faith in the replicability and dependability of the SH evaluation model. The following quote elaborates:

"We gave some of the raw data to another consultant to analyse and offer feedback on…the other consultant did check the SH data and analysis and their findings were identical to that offered by SH in their report. This is common practice here…". (Client respondent B)

6.3.5 Conclusion

When assessed as per the findings of the data collection and document review process, the SH evaluation model only partially satisfied a number of specified quality indicators, specifically those relating to the dependability audit; whilst its performance in terms of other quality indicators, such as the inclusion of study limitations and the impact thereof, was erratic and/or incomplete. However, client input indicates that SH's research is perceived as possessing a high degree of dependability and replicability. It would appear that the consultancy's performance against this criterion could easily be improved by simply applying and documenting its processes and practices in a more consistent and explicit manner. This is a key recommendation for the future practice of SH.

6.4 Confirmability

The criterion of *confirmability* addresses the post-positivist notion of 'objectivity' - and whether or not the SH evaluation model offers results that may be deemed impartial or value-neutral, and independent of the perceptions and possible bias of the evaluators. Quality indicators for the criterion *confirmability* included a confirmability audit, evidence of the use of a set of pre-defined and clearly stipulated criteria for the programme's evaluation, and reported perceptions amongst the clients of the consultancy that the evaluation process and research outputs were neutral and sufficiently inclusive of all stakeholders' input.

6.4.1 Confirmability audit:

The confirmability audit included a review of each of the following quality indicators:

6.4.1.1 Evidence of use of multiple fieldworkers / interviewers

All of the evaluations included in the meta-evaluation noted the use of three or more fieldworkers. This is an important consideration for confirmability, given that the use of multiple fieldworkers implies that no one viewpoint or approach will dominate the research process. The employment of multiple fieldworkers thus ensures that a variety of perspectives will come into play in the course of data collection, whilst any individual weaknesses in data collection will not hamper the broader research process.

However, when assessing the SH model against the evidence of use of multiple fieldworkers it was noted that such evidence was not recorded or presented in a consistent manner across all of the evaluation studies. For example, the evaluation proposal, Inception Report, Draft and Final Reports of the *Evaluation of the UNJPHT* all note the use of four teams of qualitative fieldworkers (two members per team, consisting of one SH staff member and one Zambian consultant) plus a quantitative survey team, comprised of six individuals, all

recruited in Zambia and supported by an ikapadata[46] supervisor. A *Musina Evaluation Workplan for Fieldworkers*, dated 09 July, outlines fieldwork schedules for two fieldworkers, while the final report for the *Evaluation of Save the Children UK's Response to the Situation in Musina since 2008* (2010: 11) notes that "In-depth interviews with key stakeholders were conducted in Musina by two fieldworkers and telephonic in-depth interviews were conducted with the assistance of a third fieldworker." The *Research Study of Student Risk Behaviour and Evaluation of ACES Programme for HAICU* included no direct reference to fieldworker numbers of any kind. Nevertheless, fieldworker contracts and interview / focus group schedules found amongst the project documents indicate that at least six different fieldworkers were involved in data collection on the UCT campus.

6.4.1.2 Indication that fieldworker training was conducted

Only two of the five evaluations included specific mention of fieldworker training; namely, the *Research Study of Student Risk Behaviour and Evaluation of ACES Programme for HAICU*, and the *Evaluation of the UNJPHT,* where the Inception Report (2012: 4) for the latter study notes that training will be offered by SH prior to the fieldwork period; as well as ikapadata training of the quantitative field team to ensure that the team will have "...a clear idea regarding the aims and objectives of the UNJPHT as well as of the evaluation study itself" (2012: 5). The final report for this evaluation (2012: 9) confirms that fieldworker training took place on 22 October 2012 while project documents include an agenda for fieldworker training and piloting for the evaluation, dated Monday, 22 October 2012.

One could infer that some form of fieldworker training was conducted for the PDT evaluation on the basis that a fieldworker note, detailing key information pertaining to the evaluation and the aims and objectives of the trust administrators, was sourced amongst the evaluation project documents. However, no documentary evidence of fieldworker training was sourced for the remaining two evaluations.

6.4.1.3 Evidence of the use of appropriate and unbiased questioning techniques in data collection instruments

Overall, the questioning techniques employed in the data collection instruments for the five reviewed evaluations were appropriate, clear and unbiased. For evidence of this assertion, please refer to Annexure 5, where an interview schedule for the *Research Study of Student Risk Behaviour and Evaluation of ACES Programme for HAICU* has been attached.[47] While there was some evidence of leading questions in three of the five evaluations, such occurrences were rare and did not exceed more than three to five questions amongst the total sum of all of the instruments for a particular evaluation.

Examples of leading questions uncovered include the following:

[46] ikapadata is a quantitative research team which often collaborates with SH when a mixed methods approach is required.

[47] For further input on questioning techniques for each of the reviewed evaluations, the meta-evaluation summary – attached as Annexure 1 – should be reviewed.

- *"What are the changes you have noticed as a result of using LEGO materials/education in the schools in Atteridgeville? Please specify.* (Semi-structured interview schedule for the Department of Education's subject advisors: *Outcome Assessment of the Care for Education Project: "Developing Talents through Creative Play" in Atteridgeville Township*). This question would imply that changes have taken place. However, the subsequent question in the same instrument is more appropriately phrased as "Have you noticed any changes in the teachers/educators since they have started to use LEGO materials in their teaching? Please specify."
- Another example of a leading question was found in the semi-structured interview schedule for programme funders in the *Evaluation of Save the Children UK's Response to the Situation in Musina since 2008*. This question is as follows: *"How do you think children on the move and other vulnerable children in the Vhembe district have benefitted from Save the Children's interventions?"* Once again, this form of questioning implies that there have been benefits as a result of the work of SCUK. This question could possibly be phrased as "Have you observed any changes for children on the move and other vulnerable children in the Vhembe district as a result of Save the Children's interventions? If so, what changes have you observed? If not, could you give me your opinion as to why?"

6.4.1.4 Evidence that the instruments were piloted

Project documents for three of the five evaluations indicated that piloting of specific instruments was undertaken. This includes the following forms of evidence:

- An evaluation work plan for the *Research Study of Student Risk Behaviour and Evaluation of ACES Programme for HAICU* confirms that a pilot study was conducted the week of 21 July, 2009.
- The *Proposal for the (Impact and) Outcome Assessment of the Care for Education Project: "Developing Talents through Creative Play" in Atteridgeville Township* (2010: 8) notes that the learner survey would be piloted with a small sample to "…ensure that the questions are appropriate and generate quality answers."[48]
- The *Inception Report for the Evaluation of the UNJPHT* (08 October 2012: 6) notes that the quantitative survey, 1 national level semi-structured interview, 1 district level semi-structured interview and 1 district level focus group discussion schedule would be piloted prior to entering the field. However, the *Final Report* (21 December 2012: 9) notes that the beneficiary interview schedule was piloted with three respondents and the instrument subsequently revised. There is no documented justification or explanation for this discrepancy.

Again, whilst feedback from one of the SH partners indicated an awareness of the importance of piloting of instruments prior to entering the field, the application of this

[48] However, there was no further reference to the pilot or if any indication in other project documents as to whether or not changes were made to the relevant instrument.

quality assurance mechanism is erratic both within and across the consultancy's reviewed evaluations. The following quote elaborates:

"I have certainly become more aware of the importance of piloting of instruments…over time the studies I have been involved in have confirmed this for me…like the research that we conducted in Namibia for UNICEF? This is why we advocate for including a pilot – like for the evaluation of the UNJPHT." (SH partner 3)

6.4.1.5 Evidence of quality checks of submitted transcripts

Only one of the five evaluations included evidence of the conducting of quality checks of submitted transcripts; that is, the *Proposal for the (Impact and) Outcome Assessment of the Care for Education Project: "Developing Talents through Creative Play" in Atteridgeville Township* (2010: 10) notes that "...all interviewers work under the supervision of an experienced controller who checks and validates all completed questionnaires." There was, however, no further reference to quality checks of submitted transcripts in this project's documents.

While none of the other evaluations reviewed indicated such quality assurance management, it was noted - during the interviews with each of the SH partners - that interview transcripts, submitted for each research project that the consultancy undertakes, are checked by the designated project team leader upon receipt of such transcripts. Where necessary, feedback is offered to the fieldworker responsible - either for clarification purposes or for the improvement of future data collection. This process is described by one of the consultancy partners in the quote below:

"In the LEGO outcomes assessment, I checked all transcripts as they came in from the field. These are further scrutinised when we start the data coding process. Generally, if we discover any problems with any of the transcripts, we go back to the fieldworker or interviewer responsible…I remember that we had a problem with one of the focus groups for the LEGO outcomes assessment and so I went back to the facilitator and asked questions around those issues." (SH partner 2)

6.4.1.6 Evidence of use of multiple coders for inter-coding cross checks

None of the evaluation documents reviewed as part of this meta-evaluation indicated that any form of inter-coding check had been undertaken. However, it was noted during the interviews conducted with SH partners that coding and report writing for all of the consultancy's evaluation research was generally undertaken by more than one staff member. This, it was noted, allows for comparisons of generated coding sheets – and facilitates continuous dialogue regarding the coding process and any emerging themes / observations. The following quote elaborates on this:

"Coding for the UNJPHT and LEGO evaluations was undertaken by two senior consultants. This is our general way of operating. On each of our projects we share the coding and the

report writing and this helps us to reflect on the data and each other's analysis thereof. This is supported by internal team meetings throughout the evaluation process." (SH partner 2)

6.4.1.7 Evidence of peer review of coding and data analysis

None of the evaluations included in this meta-evaluation specifically documented some form of peer review of coding and data analysis.

However, as noted above, interview data obtained from the four partners indicates that the SH partners work collaboratively as 'teams' on the various evaluation projects[49]. These team members are responsible for pre-specified tasks on each project, including data coding, data analysis and report writing. The coding sheets and report sections prepared by SH staff are thus read and commented upon by other team members throughout the study. As highlighted in the quote included in section 6.4.1.6 above, internal staff meetings are also held throughout the course of any study to enable dialogue and reflection on research outcomes and – where necessary – refinement thereof.

6.4.2 Use of a set of pre-defined and clearly stipulated criteria against which the programme / intervention will be assessed:

In terms of the documentation and use of a set of pre-defined and clearly stipulated criteria for evaluation purposes, all five of the evaluations included in this review included such a set of criteria. These were defined in the proposal for each evaluation as well as in the draft and final reports. The criteria definitions were consistent across evaluations as well as consistently applied throughout each individual evaluation.

Furthermore, all of the evaluation proposals reviewed included a list of appropriate questions - relevant to each criterion - that would be used to assess the individual programme. Thus all of the criteria utilised by SH were clearly defined and operationalised. The following forms of evidence are included to substantiate this observation:

- The *Final Report: Research Study of Student Risk Behaviour and Evaluation of ACES Programme for HAICU* (July 2009: 4) notes that the evaluation aimed to determine the effectiveness, impact and relevance of the programme for students and for ACES themselves. The findings are reported against each of these criteria, with each section commencing with a description / definition of the term. For example, in Section 3.2 of Part One (p18) it is noted: "In this section the effectiveness of the ACES programme is explored by investigating whether the programme achieved what was intended in terms of implementing the programme."
- The criteria utilised for the *Evaluation of the Ponahalo De Beers Trust* are noted in the proposal and final evaluation report. These are listed as being effectiveness, impact and sustainability. In addition to this, the documents note that the PDT

[49] The designated team for each evaluation project is noted in the project's submitted proposals. Please refer to Annexure 6, for an example thereof, taken from the *Proposal for the Evaluation of Save the Children UK's Response to the Situation in Musina since 2008 (June 2010)*.

Programme was assessed in terms of the strengths and weaknesses of the various projects (p2). The proposal includes a definition of each of the criteria, plus provides a list of appropriate questions for assessment of each of each criterion (p4). The Final Report (2012: 30) again includes a description or definition of each of the key criteria, prior to discussing the evaluation findings for each.

In addition to the above, two of the evaluation project proposals; that is, for the *Evaluation of Save the Children UK's Response to the Situation in Musina since 2008* and the *Evaluation of the UNJPHT*, explicitly note that no weighting of criteria would be utilised, but that all criteria would be of equal value in the overall programme assessment. While this was not explicitly stated in any of the other evaluations, there was no indication that any such weighting had taken place. It was not reflected in the findings narrative of any of the other final evaluation reports.

Finally, the interviews conducted with the SH partners confirm that the incorporation of the DAC criteria; namely, effectiveness, efficiency, impact, sustainability and relevance, for assessment of development intervention evaluations, is a core feature of the consultancy's evaluation model. This is illustrated by the quote recorded below:

"Underpinning the model is the use of the DAC criteria – we use these to get a good picture of what is working and what is not." (SH partner 1)

6.4.3 Evidence that clients regard the evaluation process and outputs as being sufficiently neutral and inclusive of all stakeholders' input:

Despite the high level of client engagement and involvement, all of the client respondents felt that SH had maintained a level of objectivity and neutrality, which had been reflected in the outputs of the evaluation. The following quote elaborates on this:

"I think that initially we had intended it to be an external evaluation so as to offer an objective lens on our work and what we were trying to achieve, which we could present in this way to our funders and other interested parties. But as the evaluation went on, we felt that we were sufficiently involved to make it more of a collaboration, but overall it still maintained its objectivity, which was good. It was what we wanted." (Client respondent G)

The client respondents also noted that the perspectives and opinions of all evaluation participants had been respected and fairly represented. The quotes below explain:

"I would say absolutely – we found them to be very objective, plus a broad range of stakeholder views were represented. Also the process itself, in terms of how it was carried out…SH conducted their fieldwork and spoke to relevant stakeholders on their own; they conducted their own meetings with the relevant stakeholders so their views could be honestly recorded and represented." (Client respondent F)

"Absolutely…for me this was a very strong point for SH. Also in the kind of report that they offered – I felt that they were very impartial and as close as possible to objective as one can be in this situation. They definitely included everyone's views and took that into consideration, but they were also open to the diversity of it all." (Client respondent C)

As previously noted, high levels of engagement between evaluator and evaluated are often perceived as contributing towards a certain level of co-option of the evaluator. This concern was voiced by one of the client respondents, who noted:

"I would call it a joint evaluation and yet…at the same time…this is also a challenge, not so? It is a challenge to the outcome and results of the study…too much participation and then you risk losing some of the impartiality. Or, at least, some people may perceive it that way. I was very happy with the levels of disclosure and transparency that we had with SH. There was absolute transparency here." (Client respondent C)

The feedback obtained from the consultancy's members indicates that the SH partners have a good grasp of these issues and try to address them by offering as balanced a perspective as possible in terms of the evaluation outcomes. As noted by two of the partners:

"This can be slightly tricky – do you adjust your findings because the client is unhappy with this? It is my experience that it is good to be very clear about your data here so that you are not asked to make too many changes – distribute report to as many people as possible to make sure that you get the most possible input. There are many ethical questions regarding how much to change; you know, the findings are the findings and we are clear about that. We have had some instances where this has happened – where the client may become defensive or may not understand that an evaluation is a learning process. We need to overcome that." (SH partner 3)

"Something else that we are mindful of is to find the balance in terms of doing credible research. By that I mean making the research user-friendly, but staying true to what the results are saying. Our feedback to the client at the end of the evaluation must be digestible, but still honest." (SH partner 4)

6.4.4 Conclusion

Once again, SH's performance against the quality indicators for the criterion *confirmability* is erratic. The use of a pilot study is noted in some evaluations, but not in others; and where discrepancies exist, these are not justified or explained. The same applies to the use of multiple fieldworkers and the inclusion of fieldworker training. Similarly, the inclusion of quality checks of submitted transcripts, the use of multiple coders for inter-coder cross checks and the peer review of coding and data analysis are seldom if ever explicitly outlined in evaluation project documents, yet verbal feedback from the four partners indicates that these processes are all part of the general SH procedure or evaluation model.

Client feedback, however, supports the notion that the SH evaluation model demonstrates a high level of confirmability, which is strengthened by the consultancy's consistent use and application of the DAC criteria for evaluation assessment purposes.

Cognisance needs to be taken of the fact that these evaluations have taken place over the course of the past four years and it might be argued that the consultancy's practice is becoming increasingly structured as the partners and their affiliates gain knowledge and practical experience within the development evaluation sector. However, a close scrutiny of the analytical framework will indicate that performance varies across all four years covered by this meta-evaluation with evaluations from the former years often meeting a higher level of compliance with quality indicators than those conducted only two years ago.

6.5 Meta-evaluation summary

Greene et al (1988) (cited in Cooksy and Caracelli 2009: 9) ask the pertinent question of "How many questions must be answered in the negative or how serious must an evaluator's errors be for the evaluation to 'fail'?" This is echoed by Stake (cited in Cooksy and Caracelli 2009) when he argues that it is a difficult task to arrive at an overall sense of an evaluation project's worth.

A close interrogation of the SH model reveals that, while it meets the criteria of credibility and transferability, a lack of explicitness in terms of sampling, data collection, data analysis, confirmability audit mechanisms, and effects of study limitations upon research results detracts from assessments of its dependability and confirmability. This is compounded by a lack of consistency in reporting against quality indicators. However, evidence for some of the relevant quality indicators was verbally noted by the consultancy's partners as being part of their evaluation process.

While client respondents' feedback is overwhelmingly positive, it would still appear that the current debate regarding methodological rigour is impacting upon possible perceptions of the consultancy's evaluation research outputs, leading to an awareness amongst the clients of SH of the possible 'pitfalls' of a lack of methodological explicitness and clarity, particularly when coupled with a highly participatory model. This, it might be argued, was indicated in the calls for increased levels of ownership and authority on the part of SH, particularly in terms of methodological decisions. Addressing these concerns would ensure that the consultancy's evaluation processes are seen as trustworthy and rigorous amongst a wider audience, and thus as a more 'accurate' reflection of the client organisations' efforts. This is a vitally important consideration given the resource- and funding-constrained environment within which most of these organisations are required to operate.

The following table offers an overview of the findings outlined in the previous sections.

Key Criteria and Quality Indicators	Overall Assessment
Credibility	
1.1 Evidence of member checks / reflections on research design, process, findings and recommendations.	1.1 **Meets quality indicator requirements** (based upon documented evidence, plus SH and client feedback)
1.2 Evidence of investigator and/or data source and/or methodological triangulation.	1.2 **Meets quality indicator requirements** (based upon documented evidence, plus SH feedback)
1.3 Evidence of external review of the research process and findings.	1.3 **Meets quality indicator requirements** (based upon documented evidence in three of the five evaluations, plus SH feedback)
1.4 Rationale and description of sample composition and selection (for example, characteristics of sample, basis for inclusions and exclusions, sample size and how sample allowed comparative data and / or negative case analyses to be undertaken).	1.4 **Partially meets quality indicator requirements** (based upon documented evidence, including limited input regarding sampling methods and implications thereof and erratic reporting of sampling method and rationale)
1.5 Client feedback indicates that the SH outcome and / or impact evaluation outputs are perceived as being trustworthy, credible and believable.	1.5 **Meets quality indicator requirements** (based upon client feedback)
1.6 Client feedback indicating that the SH outcome and / or impact evaluation outputs are perceived as being relevant and useful, incorporating supportive evidence; that is, examples of client utilisation of evaluation findings and recommendations.	1.6 **Meets quality indicator requirements** (based upon client feedback)
2. Transferability	
2.1 Evidence of 'thick', detailed descriptions of research context to enable readers' assessment of transferability.	**Meets quality indicator requirements** (based upon documented evidence)
2.2 Evidence that client organisations regard SH evaluation outputs to be of value in / transferable to other contexts or programmes of a similar nature.	**Meets quality indicator requirements** (based upon client feedback)
3. Dependability / Auditability	
3.1 **Dependability audit:** 3.1.1 Inclusion of detailed audit trail of data collection process and procedures thereof. 3.1.2 Inclusion of detailed audit trail of data analysis process and procedures.	 3.1.1 **Partially meets quality indictor requirements** (based upon partial and erratic documented evidence) 3.1.2 **Partially meets quality indictor requirements** (based upon limited and erratic documented evidence, but noted by SH partners and one client)
3.2 Documentation of any changes to evaluation design, including reasons and any implications thereof.	Not applicable to any of the evaluations under review
3.3 Documentation of any study limitations and their possible impact upon the evaluation.	**Partially meets quality indictor requirements** (based upon documented

	evidence, which included limitations but no clear indication of their impact upon evaluation)
3.4 Evidence that client organisations regard the SH evaluation outputs to be stable and replicable across researchers and methods.	**Meets quality indicator requirements** (based upon client feedback)
4. Confirmability	
4.1 **Confirmability audit:** 4.1.1 Evidence of use of multiple fieldworkers /interviewers. 4.1.2 Indication that fieldworker training was conducted. 4.1.3 Evidence of the use of appropriate and unbiased questioning techniques in data collection instruments. 4.1.4 Evidence that data collection instruments were piloted. 4.1.5 Indication that quality checks were conducted on submitted transcripts. 4.1.6 Indication of the use of multiple coders for inter-coder cross-checks. 4.1.7 Use of peer review of coding and data analysis.	4.1.1 **Partially meets quality indicator requirements** (use of multiple fieldworkers noted in some but not all project documents) 4.1.2 **Partially meets quality indicator requirements** (fieldworker training noted in two of the five projects' documents, plus SH fieldworker note for a third project) 4.1.3 **Meets quality indicator requirements** (based upon documented evidence) 4.1.4 **Partially meets quality indicator requirements** (noted in three of the five studies, confirmed by SH input) 4.1.5 **Partially meets quality indicator requirements** (noted in only one of the five studies, but confirmed by SH input) 4.1.6 **Partially meets quality indicator requirements** (no documented evidence, but confirmed by SH input) 4.1.7 **Partially meets quality indicator requirements** (no documented evidence, but confirmed by SH input)
4.2 Use of a set of pre-defined and clearly stipulated criteria against which the programme /intervention will be assessed.	4.2 **Meets quality indicator requirements** (based upon documented evidence)
4.3 Evidence that client organisations regard the evaluation process and outputs as being sufficiently neutral and inclusive of all stakeholders' input.	4.3 **Meets quality indicator requirements** (based upon client feedback)

Chapter 7 Discussion and recommendations

7.1 General discussion regarding findings

As previously discussed – in Section 2.2 of this project - the 'gold standard' of RCTs has re-emerged in a development sector characterised by heightened calls for accountability, a paucity of funding, and increasing demand for results- and evidence-based research. The assurance of rigour, objectivity and scientific 'certainty' offered by experimental and certain quasi-experimental methodologies is contributing towards an increased demand – and level of popularity – for such methods within the evaluation sector. However, approaching evaluations solely from a positivist or post-positivist perspective has been noted to present a number of limitations in applied developmental settings, which are complex, heterogeneous, and fluid. Thus the call from those operating within the naturalistic / constructivist paradigm for the use of more flexible, qualitative research approaches within the fields of development and evaluation studies.

Such calls have, however, been criticised on the basis that operating within the naturalistic / constructivist paradigm offers research results that are untrustworthy and which lack scientific rigour or 'worth'. As noted by Lincoln and Guba (1985, cited in De Wet and Erasmus 2005: 27) "The naturalistic inquirer soon becomes accustomed to hearing charges that naturalistic studies are undisciplined; that he or she is guilty of 'sloppy' research, engaging in 'merely subjective' observations, responding indiscriminately to the 'loudest bangs or brightest lights'. Rigor, it is asserted, is not the hallmark of naturalism." However, Lincoln, Guba, Patton, Mertens, and House (amongst many others) have robustly engaged with this debate and argue that it is, indeed, both feasible and possible to work towards rigour within a naturalistic / constructivist paradigm, and - on the basis of this research project - I concur. Using Lincoln and Guba's trustworthiness criteria as the foundation for a meta-evaluation framework, I have demonstrated that hallmarks of rigorous practice can be developed – and applied – to an evaluation research model operating within the naturalistic paradigm.

However, there appears to be a high level of disjuncture between the concept of naturalistic / qualitative research rigour and the practical application and incorporation of benchmarks or standards for rigorous practice. The lack of methodological specificity and explicitness reported against the SH evaluation model is not found within this consultancy's practice alone. Spencer et al (2003: 89) note "…recurring conclusions about the quality of reporting…" in evaluations that employ predominantly qualitative or naturalistic research methods. These authors go on to state that "In particular, it was noted that reporting of methods or 'the way the study was done' was limited or incomplete." (2003: 89) Similarly, Boulton et al (cited in Spencer et al 2003: 89) refer to the "…failure to adequately report the processes involved in analysis and interpretation…"; while Gearing (2004, cited in Ryan-Nicholls and Wills 2009) notes the tendency to state that something was done, but not *how*.

Farmer et al (2006, cited in Ryan-Nicholls and Wills 2009: 75) also assert that the richness of description in qualitative work is "…often missing from their own methodological descriptions."

This reported lack of precision and specificity in methodological reporting, across a variety of qualitative research reports, is cause for reflection – and some concern. I would argue that this lack of methodological explicitness in naturalistic / qualitative research reports is a key contributor to perceptions of a 'lack of rigour', and propose that the inclusion of – and reporting against – hallmarks of rigorous research practice will contribute towards a higher level of perceived validity - and general acceptance - of qualitative evaluation research outputs. Donaldson (cited in Donaldson and Lipsey 2006: 67) supports this argument when he argues that "A renewed emphasis on the reliance of evaluation on systematic scientific methods is especially important for overcoming negative images of evaluation as unreliable, soft, or a second-class type of investigation."

However, such 'systematic scientific methods' do not have to be linked to a specific methodological ideology. Tracy (2010: 839), for example, speaks of "universal hallmarks of quality" while Seale (1999: 466) argues that social researchers need to "…break free from the obligation to fulfil philosophical schemes through research practice…". Linking my ideas with those of these authors, and based upon my experience as a development evaluation practitioner, I advocate for a research practice which operates relatively autonomously from adherence to a specific philosophical, political or theoretical position – and the constraints or parameters within which one might make methodological decisions as a consequence of one's adopted position. Like SH, I argue for the adoption of a pragmatic approach, which would allow for informed and reflective methodological decisions on the basis of the contextual imperatives of an evaluation study. I would argue that operating solely on the basis of one's methodological paradigm is also tantamount to 'non-scientific' practice. However, whilst operating within such a pragmatic approach - and utilising quantitative, qualitative or mixed methods, as contextually required - one must not lose sight of the benefits of reflexive engagement with the hallmarks of rigorous practice.

In the current dearth of research on evaluation theory and practice, a wide variety of opportunities exist for future study and reflection on the benefits and merit of different types of evaluation practice as well as on varying forms of quality criteria frameworks, which will "…help the field move past generic and relatively abstract standards and guiding principles to more empirically supported guidance…" (Mark 2008: 115). This, it is argued, will assist with the professionalisation and improvement of evaluation practice far more than the paradigmatic contests that this sector is currently being subjected to. It is hoped that this research project has contributed in some way towards such future debates.

7.2 Recommendations

The following section outlines a number of recommendations formulated on the basis of this study's findings and conclusions. Methods that might be employed by SH as a means of improving upon its current practice include the following:

- The more consistent use of programme theory or logic models, as it is widely believed that the adoption of clear programme objectives offers a sound means of measuring and assessing such a programme's progress;[50]

- Continued use of the pragmatic approach is strongly advised; that is, the choice of method will depend upon consultation with the client organisation regarding what would constitute credible evidence for them, given the programme context and timeframe, and what might be feasible given budget and time limitations. However, methodological choice should not only be determined by such 'appropriateness of fit' but also by a heightened knowledge and ownership of the application of different methodological options, by the consultancy;

- To continue their use of subject or sector specialists to heighten the credibility of their findings and recommendations;

- The continued adoption – where possible and feasible - of a good mixed methods design[51], including quantitative and qualitative data for triangulation purposes;

- A more explicit framework for the interpretation of qualitative data, as well as higher levels of methodological input in the final evaluation report[52];

- The use of practice and methodology reflection sessions on an annual basis, coupled with capacity building of consultancy members in terms of instrument design, sampling and research methods;

- The establishment of more explicit, pre-evaluation parameters regarding the extent of the changes that can be made to the final evaluation report; and

- To overcome the prevailing model that it is the evaluation information user's value system that informs the final output, it is strongly recommended that the results of

[50] Rossi and Chen (cited in Donaldson and Lipsey 2006: 66), Bamberger and White (2007), Coryn et al (2011), Ton (2012), and Picciotto (2012) all argue that the incorporation of programme theory or Theory-Based Evaluation practices will not only lead to improvements in practice, but that it will also make such evaluations a more "rigorous and thoughtful scientific endeavour". Picciotto (2012: 226) states that while experimental and quasi-experimental methods can indicate attribution in some circumstances, theory-based and process evaluations are "…better equipped to answer how and why the observed effects have materialised."

[51] The use of mixed methods is also advocated by Campbell (cited in House 1980) who argues that qualitative data 'recontextualises' the interpretation of quantitative data and thus heightens the validity of the interpretive process.

[52] This may be attached as an appendix to the main report to meet any report length requirements of the client organisation.

the impact evaluation / outcome assessment be shared with the target groups and beneficiaries as well (this will also heighten the perceived credibility of the evaluation research findings).

In general, it is advocated that SH engage with the contemporary debate around hallmarks of rigorous practice more extensively. This does not necessarily mean that the consultancy adopt a more post-positivist approach or quantitative methodologies. It is, however, recommended that they investigate possible means of increasing the perceived credibility of their qualitative and mixed methods design - and the possible impact that such decisions may have upon the research outcomes. This is deemed to be particularly relevant in light of high levels of client involvement, which is often perceived as detracting from an accurate and honest evaluation or programme appraisal. I argue that such an approach will give SH a greater 'voice' and level of authority, whilst ensuring that their fundamental, utilisation-focused and participatory approach remains adaptable, flexible, practical, and cognisant of their clients' needs.

Chapter 8 Conclusion

The aims of this research project were: Firstly, to interrogate whether or not a particular evaluation approach or model could deliver a rigorous or 'quality' evaluation of development interventions' outcomes and impact, utilising the criteria of credibility, transferability, dependability, and confirmability as hallmarks of rigorous practice; secondly, to yield specific insights or findings to enable critical reflection upon – and possible improvement of – this specific evaluation approach or model, thus encouraging the developers and implementers thereof to think more systematically and critically about the work that they currently undertake; and thirdly, to contribute towards the current paradigmatic debate regarding the use of quality standards and criteria as a means of assessing – and informing – naturalistic evaluation practice.

A meta-evaluation of five outcome assessments and impact evaluations, conducted over the course of the past four years, was conducted. This meta-evaluation included both primary and secondary data collection; that is, interviews with both SH partners as well as consultancy clients, coupled with an extensive project document review. An Excel spreadsheet meta-evaluation summary was compiled to facilitate data analysis together with primary data coding and analysis utilising NVivo 10 software.

In answering the central research question; that is, *Can the outcomes and impact of development interventions be assessed with an appropriate level of rigour by using SH's naturalistic evaluation model?* I would argue – on the basis of the input obtained during this study process – that SH does, indeed, produce a credible and transferable measurement of the outcomes and impact of development interventions. This is achieved via their use of a highly collaborative, relevant, flexible and practical approach; their inclusion of a wide range of stakeholders; and their considerable experience in the fields of development, evaluation and qualitative data collection methods. This allows SH to deliver 'rich', perceptive and in-depth information that charts and documents intervention outcomes from the perspective of those experiencing it. However, the model falls short in term of its dependability and confirmability due to a lack of methodological clarity, consistency and transparency.

Of course, like most aspects of the development evaluation sector, the findings and conclusions of this research project are open to debate, specifically regarding the use and application of a set of quality criteria for a naturalistic evaluation approach. However, as noted in this research report, the consideration and application of hallmarks of rigorous practice will undoubtedly facilitate a level of reflection and improvement in evaluation practice, whilst ensuring that the generated evaluation results are not simply dismissed by critics as being unsound, untrustworthy and of poor quality.

I would like to conclude with a quote from Patton (2002: 189) who stated that "Findings have a very short half-life…they deteriorate very quickly as the world changes rapidly. Specific findings typically have a small window of relevance. In contrast to this, learning to

think and act evaluatively can have an on-going impact." If this were the sole yardstick by which to interrogate the SH model, I would deem it to be a highly successful evaluation process. As noted by one of the consultancy's clients:

"I think that the process was even more beneficial than the outcomes. There were some key learnings that took place there…this was a very beneficial process for me, and for us…, to undergo." (Client respondent D)

Sources Consulted

African Evaluation Association: **www.afrea.org**. Accessed 12/12/2012; 13/12/2012.

American Evaluation Association: **www.eval/org/EvaluationDocuments/progeval**. Accessed 8/11/2012; 5/12/2012; 12/12/2012; 13/12/2012.

Babbie, E. and Mouton, J. 2007. *The Practice of Social Research*. Oxford: Oxford University Press.

Bamberger, M. and Rugh, P. 2008. *A Framework for assessing the quality, conclusion validity and utility of evaluations.* Paper presented at the American Evaluation Association Conference, Baltimore, 7-10 November 2007.

Bamberger, M. and White, H. 2007. Using Strong Evaluation Designs in Developing Countries: Experience and Challenges, in *Journal of MultiDisciplinary Evaluation*, 4 (8): 58-73.

Barbour, R.S. 2001. Checklists for improving rigour in qualitative research: a case of the tail wagging the dog? in *British Medical Journal*, Vol 322: 1115-1117.

Bhola, H.S. 2000. A Discourse on Impact Evaluation: A Model and its Application to a Literacy Intervention in Ghana, in *Evaluation* 6 (2): 161 – 178.

Bickman, L. and Reich, S.M. 2009. Randomised Controlled Trials: A Gold Standard with Feet of Clay? in *What Counts as Credible Evidence in Applied Research and Evaluation Practice?* Donaldson, SI, Christie, CA and Mark, MM (Eds). Thousand Oaks, California: Sage.

Center for Disease Control and Prevention: **http://www.cdc.gov/eval/standards/index**. Accessed 12/12/2012; 13/12/2012.

Chianca, T. 2008. The OECD / DAC Criteria for International Development Evaluations: An Assessment and Ideas for Improvement, in *Journal of MultiDisciplinary Evaluation*, 5 (9): 41-51.

Christie, C.A. and Fleischer, D. 2009. Social Inquiry Paradigms as a Frame for the Debate on Credible Evidence, in *What Counts as Credible Evidence in Applied Research and Evaluation Practice?* Donaldson, SI, Christie, CA and Mark, MM (Eds). Thousand Oaks, California: Sage.

Colorado State University Glossary of Terms: **www.colostate.edu**. Accessed 25/06/2013.

Cook, T.D. 2006. Describing what is special about the role of experiments in contemporary education research: Putting the 'gold standard' rhetoric into perspective, in *Journal of MultiDisciplinary Evaluation*, 6 (): 1-7.

Cooksy, L.J. and Caracelli, V.J. 2009. Metaevaluation in Practice: Selection and Application of Criteria, in *Journal of MultiDisciplinary Evaluation*, 6 (11): 1-15.

Cooksy, L.J. and Mark, M.M. 2012. Influences on Evaluation Quality, in *American Journal of Evaluation* 33 (1): 79 – 87.

Coryn, C.L.S. 2007. The Holy Trinity of Methodological Rigour: A Skeptical View, in *Journal of MultiDisciplinary Evaluation*, 4 (7): 26-31.

Coryn, C.L.S., Noakes, L.A., Westine, C.D., and Schroter, D.C. 2011. A Systematic review of Theory-Driven Evaluation Practice from 1990 – 2009, in *American Journal of Evaluation* 32 (2): 199 – 226.

Davey, J.W., Gugiu, P.C. and Coryn, C.L.S. 2010. Quantitative Methods for Estimating the Reliability of Qualitative Data, in *Journal of MultiDisciplinary Evaluation,* 6 (13): 140-162.

Davidson, E.J. 2005. *Evaluation Methodology Basics: The Nuts and Bolts of Sound Evaluation*. Thousand Oaks, California: Sage.

Development Assistance Committee Network on Development Evaluation. 2010. *Evaluating Development Cooperation; Summary of Key Norms and Standards (2nd Edition)*. OECD / DAC Network on Development Evaluation.

De Wet, J. and Erasmus, Z. 2005. Towards rigour in qualitative analysis, in *Qualitative Research Journal*, 5 (1): 27-40.

Dillman, L.M. 2013. Comparing evaluation activities across multiple theories of practice, in *Evaluation and Programme Planning*, 38 (June): 53-60.

Dixon-Woods, M., Shaw, R.L. and Agarwal, S. et al. 2004. The problem of appraising qualitative research, in *Quality Safety Health Care* 13: 223-225.

Donaldson, S.I. and Christie, C.A. 2005. Determining Causality in Program Evaluation and Applied Research: Should Experimental Evidence Be the Gold Standard? in *Journal of Multidisciplinary Evaluation*, 2 (3): 60-77.

Donaldson, S.I. and Lipsey, M.W. 2006. Roles for Theory in Contemporary Evaluation Practice: Developing Practical Knowledge, in *The Handbook of Evaluation: Policies, programs and practices*. Shaw, I, Greene, JC and Mark, MM (Eds). London: Sage Publishers.

Donaldson, S.I. 2009. Epilogue: A Practitioner's Guide for Gathering Credible Evidence in the Evidence-Based Global Society, in *What Counts as Credible Evidence in Applied Research and Evaluation Practice?* Donaldson, SI, Christie, CA and Mark, MM (Eds). Thousand Oaks, California: Sage.

Durrheim, K. and Wassenaar, D. 1999. Putting design into practice: writing and evaluating research proposals, in *Research in Practice: Applied Methods for the Social Sciences* (Eds: Terre Blanche, M. and Durrheim, K).Cape Town: University of Cape Town Press.

Eggers, H.W. 2009. Comments and Proposals Concerning Chianca's 'The OECD / DAC Criteria for International Development Evaluations: An Assessment and Ideas for Improvement', in *Journal of MultiDisciplinary Evaluation* 6 (11): 116 – 124.

European Evaluation Society. 2007. Statement regarding RCTs: **www.europeanevaluation.org/download/?id** = 1969403. Accessed 8/11/2012; 5/12/2012.

Gillham, B. 2000. *Case Study Research Methods*. London: Continuum.

Gomm, R. 2008. *Social Research Methodology: A Critical Introduction* (2nd Edition). Hampshire: Palgrave MacMillan.

Gregory, A. 2000. Problematizing Participation: A Critical Review of Approaches to Participation in Evaluation Theory, in *Evaluation* 6 (2), pp 179 – 199.

Guba, E.G. and Lincoln, Y.S. 2001. Guidelines and checklist for constructivist (a.k.a Fourth Generation) evaluation. Source: **www.wmich.edu/evalctr/checklists**. Accessed 12/12/2012; 13/12/2012.

Guba, E.G. and Lincoln, Y.S. 1989. *Fourth Generation Evaluation*. Newbury Park, California: Sage.

Hedler, H.C. and Gibram, N. 2009. The Contribution of Metaevaluation to Program Evaluation: Proposition of a Model. In *Journal of MultiDisciplinary Evaluation*, 6 (12): 210 – 223.

Henry, G.T. 2009. When Getting It Right Matters: The Case for High-Quality Policy and Program Impact Evaluations, in *What Counts as Credible Evidence in Applied Research and Evaluation Practice?* Donaldson, SI, Christie, CA and Mark, MM (Eds). Thousand Oaks, California: Sage.

Holden, D.J. and Zimmerman, M.A. (Editors). 2009. *A Practical Guide to Program Evaluation Planning.* Thousand Oaks, California: Sage Publications.

House, E.R. 1980. *Evaluating with Validity*. Thousand Oaks, California: Sage.

Joint Committee on Standards for Educational Evaluation. **www.jcsee.org/program-evaluation-standards**. Accessed 12/12/2012; 13/12/2012.

Kelly, K. 1999. Calling it a day: reaching conclusions in qualitative research, in *Research in Practice: Applied Methods for the Social Sciences* (Eds: Terre Blanche, M. and Durrheim, K).Cape Town: University of Cape Town Press.

Kirk, J. and Miller, M. 1986. *Reliability and validity in qualitative research*. Beverly Hills, California: Sage.

Krefting, L. 1991. Rigor in Qualitative Research: The Assessment of Trustworthiness, in *The American Journal of Occupational Therapy*, 45 (3): 214-222.

Kusters, C. with van Vugt, S., Wigboldus, S., Williams, B. and Woodhill, J. 2011. *Making evaluations matter: A practical guide for evaluators.* Centre for Development Innovation, Wageningen (The Netherlands): Wageningen University and Research Centre.

Le Compte, M. and Goetz, J. 1982. Problems of Reliability and Validity in Ethnographic Research, in *Review of Educational Research*, 52 (1): 31-60.

Lincoln, Y.S. and Guba, E.G. 1985. *Naturalistic Inquiry*. Beverly Hills, California: Sage.

Luskin, R.J.C. and Ho, T. 2013. Comparing the intended consequences of three theories of evaluation, in *Evaluation and Program Planning*, 38 (June): 61-66.

Mark, M.M. 2008. Building a Better Evidence Base for Evaluation Theory, in *Fundamental Issues in Evaluation*. N.L. Smith and P.R. Brandon (Eds). New York: Guilford Press.

Mertens, D.M. 2009. *Transformative Research and Evaluation*. New York: The Guilford Press.

Miles, M.B. and Huberman, A.M. 1984. *Qualitative Data Analysis: A sourcebook of new methods*. Beverly Hills, California: Sage.

Miles, M.B. and Huberman, A.M. 1994. *Qualitative Data Analysis: An expanded sourcebook* (2nd Edition). Thousand Oaks, California: Sage.

Miller, R.L. 2010. Developing Standards for Empirical Examinations of Evaluation Theory, in *American Journal of Evaluation* 31 (3): 390 – 399.

Morra Imas, L.G. and Rist, R.C. 2009. *The Road to Results: Designing and Conducting Effective Development Evaluations*. Washington: The World Bank.

Morse, J.M., Barrett, M., Mayan, M., Olson, K. and Spiers, J. 2002. Verification Strategies for Establishing Reliability and Validity in Qualitative Research, in International Journal of Qualitative Methods 1 (2): 13-22.

Patton, M.Q. 2008. Advocacy Impact Evaluation, in *Journal of MultiDisciplinary Evaluation* 5 (9): 1-10.

Patton, M.Q. 2003. Qualitative Evaluation Checklist – Evaluation Checklists Project. Source: **www.wmich.edu/evalctr/checklists**. Accessed 12/12/2012; 13/12/2012.

Patton, M.Q. 2001. *Qualitative Research and Evaluation Methods* (3rd Edition). Thousand Oaks, California: Sage.

Patton, M.Q. 1978. *Utilization-Focused Evaluation*. Beverly Hills, California: Sage.

Picciotto, R. 2012. Experimentalism and development evaluation: Will the bubble burst? In *Evaluation* 18 (2): 213 – 229.

Pike Hall, S. 1995. *Exploring the Authenticity Criteria of Fourth Generation Evaluation*. Paper presented to American Evaluation Association, Drake University, Vancouver, Canada. Source: **www.educ.drake.edu/hall/Naturalisticeval/4thGenEvalAuthentic**.

Posavac, E.J. and Carey, R.G. 2007. *Program Evaluation: Methods and Case Studies* (7th Edition). New Jersey: Pearson Prentice Hall.

Rebien, C.C. 1996. Participatory Evaluation of Development Assistance: Dealing with Power and Facilitative Learning, in *Evaluation* 2 (2): 11-172.

Reichardt, C.S. 2011. Evaluating Methods for Estimating Program Effects, in *American Journal of Evaluation*, 32 (2): 246-272.

Rogers, P.J. 2008. Using programme theory to evaluation complicated and complex aspects of interventions, in *Evaluation* 14 (1): 29-48.

Rolfe, G. 2006. Validity, trustworthiness and rigour: quality and the idea of qualitative research, in *Journal of Advanced Nursing*, 53 (3): 304-310.

Rubin, A. and Babbie, E. 2011. *Research Methods for Social Work*. Belmont, California: Brooks, Cole and Cengage Learning.

Ryan, G.W. 2009. What are the Standards of Rigour for Qualitative Research? Source: **www.wjh.harvard.edu/nsfqual/Ryan**. Accessed 25/06/2013.

Ryan-Nicholls, K.D. and Will, C.I. 2009. Rigour in qualitative research: mechanisms for control, in *Nurse Researcher*, 16 (3): 70-85.

Salmen, L.F. 2002. *Beneficiary Assessment: An Approach Described* (Paper Number 10). Social Development Department : World Bank.

Sandelowski, M. 1993. Rigor or rigor mortis: The problem of rigor in qualitative research revisited, in *Advanced Nursing Science*; 16 (2): 1-8.

Sanders, J.R. and Nafziger, D.N. 2011. A Basis for Determining the Adequacy of Evaluation Designs (originally published as Paper #7, Occasional Paper Series, April 1976), in *Journal of MultiDisciplinary Evaluation*, 7 (15): 44-78.

Sanderson, I. 2002. Evaluation, Policy Learning and Evidence-based Policy Making, in *Public Administration*, 80 (1): 1 – 22.

Saunders, N. 2011. Capturing effects of interventions, policies and programmes in the European Context: A social practice perspective, in *Evaluation* 17 (1): 89 – 102.

Scriven, M. 2008. A Summative Evaluation of RCT Methodology: An Alternative Approach to Causal Research, in *Journal of MultiDisciplinary Evaluation*, 5 (9): 11-24.

Scriven, M. 2007. Key evaluation checklist. Source: www. wmich.edu/evalctr/checklists. Accessed 12/12/2012; 13/12/2012.

Seale, C. 1999. Quality in Qualitative Research, in *Qualitative Inquiry* 5 (4): 465-478.

Southern Hemisphere website: **www.southernhemisphere.co.za**. Accessed 12/12/2012 and 14/04/2013.

Spencer, L; Ritchie, J; Lewis, J and Dillon, L. 2003. *Quality in Qualitative Evaluation: A framework for assessing research evidence*. National Centre for Social Research report, produced for the Cabinet Office, United Kingdom.

Stame, N. 2010. What doesn't work? Three Failures, Many Answers, in *Evaluation* 16 (4): 371 – 387.

Stufflebeam, D.L. 2011. Meta-Evaluation (originally published as Paper #3, Occasional Paper Series, December 1974), in *Journal of MultiDisciplinary Evaluation*, 7 (15): 99-158.

Stufflebeam, D.L. 1999. Program Evaluation Models Meta-evaluation Checklist – based on The Program Evaluation Standards. Source: **www.wmich.edu/evalctr/checklists**. Accessed 12/12/2012; 13/12/2012.

Ton, G. 2012. The mixing of methods: A three-step process for improving rigor in impact evaluations, in *Evaluation* 18 (1): 5 – 25.

Tourmen, C. 2009. Evaluators' Decision-Making: The relationship between theory, practice and experience, in *American Journal of Evaluation*, 30 (1): 7 – 30.

Townsend, K. and Burgess, J. (Editors). 2009. *Method in the Madness: Research stories you won't read in textbooks*. Oxford: Chandos Publishing.

Tracy, S.J. 2010. Qualitative Quality: Eight "Big-Tent" Criteria for Excellent Qualitative Research, in *Qualitative Inquiry*, 16 (10): 837-851.

Trochim, W.M.K. 2001. *The Research Methods Knowledge Base* (Second Edition). Cincinnati: Atomic Dog.

Unrau, Y.A., Gabor, P.A. and Grinnell, R.M. 2007. *Evaluation in Social Work: The Art and Science of Practice* (4th Edition). New York: Oxford University Press.

Wallace, T.L. 2011. An argument-based approach to validity in evaluation, in *Evaluation* 17 (3): 233 – 246.

White, H. 2009. Theory-based impact evaluation: principles and practice, in *Journal of Development Effectiveness*, 1 (3): 271 – 284.

White, H. 2010. A Contribution to Current Debates in Impact Evaluation, in *Evaluation* 16 (2): 153 – 164.